Pruning the image set for appearance based robot localization

O. Booij Z. Zivkovic B. Kröse

Intelligent Autonomous Systems Group
University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{obooij, zivkovic, krose}@science.uva.nl

Keywords: Robot localization, appearance based, topological maps, graph theory

Abstract

In appearance based robot localization a new image is matched with every image in the database. In this paper we describe how to reduce the number of images in this database with minimal loss of information and thereby increasing the efficiency of localization significantly. First we build a low level representation that consists of a graph in which relations between images are represented. We use a metric based on visual landmarks (SIFT features) and geometrical constraints. This graph is then pruned using the Connected Dominating Set algorithm. The method is applied on real data and evaluated by correlating new images with images in the Connected Dominating Set.

1 Introduction

To effectively navigate from one place to another a mobile robot needs an internal representation, or model, of its environment. Before it can start to plan a path, the robot must first determine its present location in this internal representation, the so called localization task. This is accomplished by sensing its surroundings and finding that spot in the model that has the highest possibility of producing the sensed information. The accuracy of this procedure does not only depend on the type of sensors and the matching algorithm but also on the type of representation of the environment.

There are, roughly stated, two types of techniques of building a model using sensor data. The traditional approach is to try to maintain a geometric model of the task space of the robot. A well studied example of this type of model is the occupancy grid, which resembles a discretized 2D or 3D representation of the environment [17]. The advantage of such a metric model is that there is a one-to-one correspondence between a position in the model and a position in the environment, which makes it relatively easy to reason about tasks in the environment, such as finding optimal paths. However, for simple low level tasks that involve the processing of new sensor readings, such as localization, this method is less practical. To find the position in the geometric model that could have produced the sensor data is a difficult task, even if it has already visited the exact same position.

A relatively recent approach is to not explicitly model the environment, but to model the way that environment appears to the robot, hence the name: appearance based modeling [10]. In appearance based approaches we explicitly memorize the appearance of the surroundings for every admissible location. The model, which then consists of a map of sensor readings obtained at known locations, is far from being a one-to-one model of the environment, but more a self-centered model of how the environment appears to the robot.

Our robot uses such an appearance based approach by memorizing a set of images taken from the environment [20]. We will however not make use of the geometrical position from where the images were taken, except for visualizing the map. The robot is thus only provided with an unordered set of images taken from different locations in the environment. The images in the set are then correlated pairwise to form a graph, which could be considered to be a topological map of the environment. This correlation is performed by matching the local landmarks after they are

detected in every image as described in [11]. By demanding the possibility of a 3D reconstruction of the matched landmarks, the method is made robust to prevent false matches, that could result from perceptual aliasing.

Localizing oneself in such a topological map implies finding the node in the graph that resembles the current position. For appearance based approaches this is achieved by taking a new image and finding the best matching image in the image set. In our research we make use of an omnidirectional camera, thus every image uniquely describes a certain location in the environment. When using an ordinary camera more images can be taken on the same location describing not only the position but also the viewing direction of the robot.

If the number of images in the database is very large the matching process poses a problem, because many images should be matched, costing a lot of computational resources. Thus, it would be advantageous if the number of images in the database were reduced. Especially images that are very similar to other images can be discarded without losing much information. The similarity among the images is already present in the graph structure of the topological map, which was constructed by matching them. In this paper we describe a pruning technique originating from graph theory, called the Connected Dominating Set problem, which will reduce the graph considerably. In this way we end up with a much smaller amount of images, which still gives a good representation of the environment.

The rest of this paper is organized as follows. Section 2 starts with an overview of other approaches of reducing the number of features in appearance based models and related work. Then, in Section 3 we describe how the topological graph is build using the images. Section 4 presents the graph theoretical pruning technique, which will be used to reduce the topological map. Section 5 gives some practical details about the two-view geometric constraints we use to build our maps. We test the proposed methods on an image set obtained by our robot and report the results in Section 6. And finally in Section 7 we make some conclusions and give directions for future work.

2 Related work, reducing the number of features in appearance based systems

All applications of appearance based methods have to deal with the same problem of selecting which observations to store in its data base. This does not only account for robot localization systems, but in general for appearance based vision tasks, such as object recognition [5]. Nevertheless, in most research a crude method is used to create a database of images. A common technique is to make a grid of images, acquiring images proportional to the size of the surface of the environment, as in [16, 9]. A comparable method is to take one image per unit time, while driving around. In [19] for example images are taken at the rate of 1 Hz. If computational resources are not a problem or the task space is small, these simple approaches can be sufficient.

In general however the total amount of observed data should be reduced. But instead of removing images it is also possible to retain only the most salient features from the image set. In [13] the minimal number of features per region in the environment is calculated so every image in that region can still be localized. Features that appear in all images of a region are thus regarded as good features. Similar methods do not calculate the optimal minimum but track features in consecutive images, while calculating there distinctiveness [16, 15].

Another technique that is related to our work is the visibility graph which can be seen as a specific task. The question is how to compute the minimal subset of locations in a given geometrical map, so every point in the map is visible from at least one of these locations[7]. The work presented here is similar while the minimal subset of image-locations needs to be found from where all other images are visible. However in our case the geometrical layout of the environment is unknown. Furthermore the research of visibility graphs is solely based on range detectors, such as sonars, while we use a vision based system.

3 Topological Map From Images Using Appearance and Geometrical Constraints

A general definition of a topological map is that it is a graph-like representation of space. The nodes of the graph represent positions and poses in space and the edges encode how to navigate from one node to the other [12]. The nodes and the edges are usually enriched with some local metric information.

In this paper, as it is typical in the appearance based approaches, each node presents a location and it is described by an image taken at that location. We are using SIFT features [11] as the automatically detected landmarks. Therefore an image can be summarized by the landmark positions and descriptions of their local appear-

ance. We define that there is an edge between two nodes in the graph if it is possible to perform 3D reconstruction of the local space from the two corresponding images. The algorithm we were using for the 3D reconstruction is described in Section 5. As the result from n images we get a graph that is described with a set of n nodes V and a symmetric matrix S called the 'similarity matrix'. For each pair of nodes i,$j\epsilon[1, ..., n]$ the value of the element S_{ij} from S defines the similarity of the nodes. In our case this is equal to 1 if there is an edge between the nodes and 0 if there is no edge. An example of such graph that we obtained from a real data set is given in figure 4.

For localization and navigation the robot could use the same computer vision algorithm as the one that was used to define the edges of the graph (V, S) [1]. An edge in the graph denotes that the 3D reconstruction is possible between the images that correspond to the nodes. This also means that if the robot is at one node it can determine the relative location of the other node. Therefore, if there are no obstacles in between, the robot could navigate from one node to the other (for example as in [2]). If there are obstacles, we could rely, for example, on some lower level algorithms for obstacle avoidance that is using range sensors. Furthermore, additional information can be associated with the edges of the graph. For example, if we reconstruct the metric positions of the nodes (using the images or we measure them in some other way), we could also associate the Euclidean distance between the nodes with each edge. This could be used to navigate through the graph. However, this is beyond the scope of this paper.

The graph will contain, in a natural way, the information about how the space in an indoor environment is separated by the walls and other barriers. Images from a convex space, for example a room, will have many connection between them and just a few connections to some images from another space, for example a corridor, that is connected with the room via a narrow passage, for example a door. The connectedness of the nodes represents the level of overlap of information among the connected images. So a place in the graph where the nodes are highly connected indicate that the environment viewed from those nodes is very similar, while a group of nodes with very few connections indicates that different views are highly distinctive.

The objective of this research is to end up with a minimal set of images that still gives a good representation of the environment. All images that add very little information to the dataset because they are very similar to images already in the set should be left out. Using the information contained in the graph structure we can accomplish just this. We state that a node that can be removed without disconnecting another node or group of nodes, can be left out of the final graph of nodes. It is simple to see that this is valid: if the robot is on the spot where the removed image was taken, it can take a new image and estimate its location in the topological map with at least one of the images in the set with which it used to be connected. Finding the maximum number of nodes, and thus images, that can be removed is not a trivial problem. In the next section we will explain the method we use to find an approximate solution.

4 Connected Dominating Graph

The problem we have to solve, is determining which node of the graph, defined in Section 3, can be removed, without causing other nodes to become unreachable. This problem is a known problem in graph theory called the Connected Dominating Graph problem (CDS) and is encountered in a lot of other domains, such as radio-broadcasting and computer-networking. In this section we will give an exact definition of CDS and describe a method to find an approximate solution.

4.1 Definition

For a connected graph $G = (V, E)$, where V denotes the set of nodes of the graph and E the edges between the nodes, a Dominating Graph $G' = (V', E')$ is defined as follows. The set of nodes in the CDS V' is a proper subset of the original set V, such that every node u in the original set V is either in the Dominating Set V' or is neighboring a node in V':

$$\forall u \in V : u \in V' \vee (v \in V' \wedge (u, v) \in E) \qquad (1)$$

For clarity, a CDS poses no restrictions on the set of edges E', except that it is a subset of the edges in the original graph. In case of a *Connected* Dominating Graph the subgraph G' is connected. The problem now is to find a connected subgraph with the minimal number of nodes. This task is however known to be NP-complete, but fortunately there are some algorithms that can find a good approximation in polynomial time [6].

Most of these algorithms will first remove edges to make a spanning tree with as many leaves as possible and then remove all the leaves resulting in a smaller tree. The CDS-problem however does not imply anything about the number of edges and does not have to be a tree. In our case it

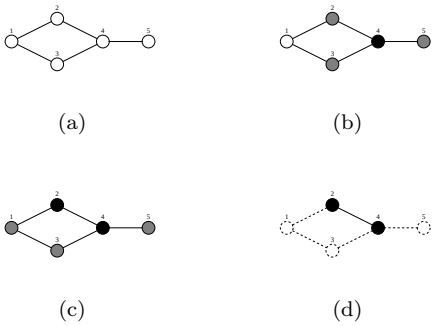

Figure 1: A simple example describing the approximation algorithm. (a) A graph of five nodes. (b) Node 4 has the most neighbors, i.e. 3, and is therefore colored black and its neighbors gray. (c) In the next step node 2 is colored black because it is connected to the only white node left (node 3 could also be chosen) (d) No white nodes are left so we have made a Dominating Graph consisting of two nodes.

is best to conserve all edges between the remaining nodes in the CDS, because they indicate a path between the two image-locations. So after the tree is pruned of his leaves, the original edges should be restored where possible.

4.2 Approximation algorithm

A couple of approximation algorithms for the CDS problem are given by Guha in [6]. We use one of these algorithms that produces a Connected Dominating Set and can be implemented using polynomial time in the order of the number of nodes. This iterative algorithm can be explained as follows, see also figure 1:

1. First color every node of the graph white (figure 1(a)).
2. Initially choose a white node with the biggest number of neighbors.
3. Color this node black and color all white neighboring nodes gray (figure 1(b)).
4. Choose a gray node that has the most edges leading to white nodes (figure 1(c)).
5. Goto 3 until there is no white node left.
6. The black nodes now compose the Connected Dominating Set (figure 1(d)).

By connecting the found nodes with edges that were in the original graph, we found a Connected Dominating Graph.

5 Visual Landmarks and Geometric Constraints

A distinctive point in an image with a well defined position in the image is called an interest-point, for example a corner, T-junction, a white dot on black background etc. Such points are often used in the computer vision community as automatically detected landmarks. We use here the SIFT feature detector [11]. The SIFT feature detector extracts also the scale of the feature point and describes the local neighborhood of the point by a 128-element rotation and scale invariant vector. This vector descriptor is also robust to some light changes.

5.1 Matching Landmarks

Visual landmarks are used often in robotics for navigation [15, 13, 16]. It is possible to reconstruct both the camera images and the 3D positions of the landmarks by matching (or tracking) landmarks through images. On-line simultaneous localization and reconstruction of landmark positions was presented in [4] but currently only for small scale environments.

In this paper we consider the general case when we start with a set of unordered images of the environment. This is similar to [14]. In practice we usually have some information about ordering of the images (a movie as in [4]) or some other sensor readings (odometry for example). This extra information should be used then.

Most 3D reconstruction algorithms [8] start with finding similar landmarks in pairs of images. When two images are consecutive frames of an image sequence we could track the landmarks from one image to the other [4]. However, it is much more difficult to find matching landmarks in an unordered set of images. Firstly, we need to check all the pairs of images which is computationally expensive. Secondly, there are no additional constraints as is generally the case in an image sequence. Therefore we might expect to have many false matches.

In this paper we use an approach similar to [14]. First we check if there are many similar landmarks within each image. Such landmarks could potentially lead to false matches. We discard the landmarks that have 6 or more similar landmarks. The landmarks are similar if the Euclidean distance between them is less than 10% of the dynamic range (elements of the SIFT descriptor vector have values between 0 and 255).

Further, for a landmark from one image we find the best and the second best matching landmark from the second image. The goodness of the

match is defined by the Euclidean distance between the landmark descriptors. If the goodness of the second best match is less than 0.8 of the best one it means that the match is very distinctive. According to the experiments in [11] this typically discards 95% of the false matches and less than 5% of the good ones. This is repeated for each pair of images and it is computationally expensive. Fast approximate methods were discussed in [11]. Since our data sets were not very big we performed the full extensive search.

5.2 Geometric Constraints

We first find the possible matches for each pair of images from our data set as described in the previous section. Let there be N matching landmark points between images m and l. The 2D image positions of the points in the m-th image in homogeneous coordinates are denoted as $\{\vec{p}_1^m, ..., \vec{p}_N^m\}$. The corresponding points in the l-th image are $\{\vec{p}_1^l, ..., \vec{p}_N^l\}$. If the i-th point belongs to the static scene, then, for a projective camera, the positions are related by:

$$(\vec{p}_i^m)^T F \vec{p}_i^l = 0 \qquad (2)$$

where the matrix F is also known as the 'fundamental matrix'. Estimating F is an initial step for 3D space reconstruction from images.

The approach [14] leads to many initial false matches. Standard robust M-estimators can deal with a certain amount of outliers. The robust algorithm called RANSAC is usually used [8] if there are more outliers. It was shown [18] that a combination that performs the best is when the RANSAC is used first and then the M-estimator. Instead of following the approach of [8] completely we use only the distinctive matches as in [11] that discards many false matches. In our experiments we observed that there were still enough good matches remaining. We used here the standard 8-point algorithm [8] which requires at least 8 matching points. With such small number of false matches it is possible to use the robust M-estimator directly. We used Huber M-estimator here.

For the whole data set we calculate the global standard deviation σ_{global} for the points. The σ_{global} is estimated robustly using maximum absolute difference estimate. This global standard deviation is used to decide when the fundamental matrix is properly calculated. The whole procedure goes then as follows:

- extract SIFT landmarks from all images
- discard self similar landmarks within each image

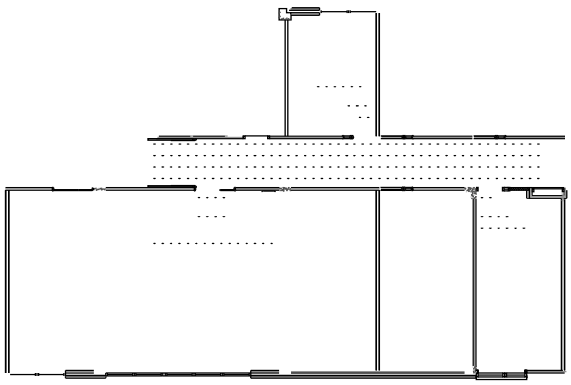

Figure 2: Bird's eye view of the environment containing a corridor and three rooms. The dots represent the image-locations.

- find distinctive matches between pairs of images
- if there are more than 8 matches:
  - estimate the fundamental matrix using M estimator (and eventually RANSAC)
  - discard the matches that deviate more than $2.5\sigma_{\text{global}}$
  - if there are still more than 8 matches then there is an edge in the graph

6 Experiments

For the experiments we use a set of 234 omnidirectional images taken by a Nomad robot using a hyperbolic camera in an office-environment consisting of a corridor with three adjacent rooms, see figure 2. The circular images are first converted to make corrected perspective panoramic images [3]. Then in all images the SIFT features are detected and matched among all image pairs as explained in Sections 3 and 5. In figure 3 an example of a matching image pair is shown with a considerable amount of occlusion and in figure 4 the resulting graph is shown using the image locations as the positions of the nodes. As can be seen from the denseness of the edges in the figure, a lot of matches were found among the image set, especially in the corridor. Because of the repetitive structure of the corridor and similarities inside the rooms, some of the matches could be the result of perceptual aliasing. However, due to the 3D reconstruction constraint, no false matches were produced.

The high level of connectedness of the graph indicates that the images in the set are highly

Figure 3: A matching image pair. The lines indicate matching landmarks between the two images.

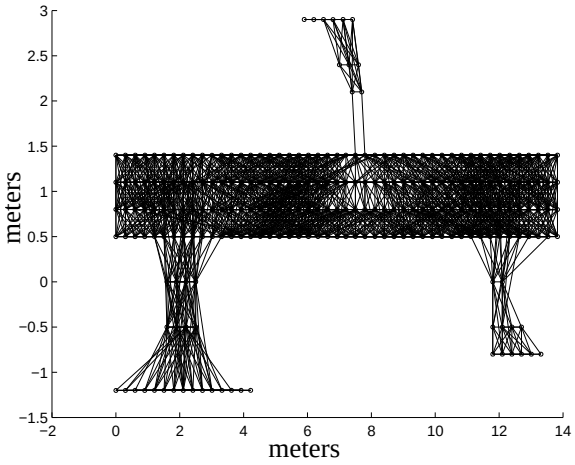

Figure 4: Graph representing the pair-wise correlation between the images in our database.

correlated. This already implies that a number of images could be left out without reducing the information in the image set. In the following sections the minimal set of images will be computed using the Connected Dominating Graph approximation algorithm and the quality of this reduced dataset will be tested by matching new images.

6.1 Computing the reduced set of images

The approximate CDS algorithm as described in Section 4 is applied to the graph obtained by the image-correlation method (see figure 4). The resulting subgraph is shown in figure 5. For clarity it must be reminded that the CDS algorithm does not use the positional information of the nodes which is used to visualize the graph. The graph is solely based on information in the images themselves.

As can be seen the graph is strongly reduced; 209 images are removed, while only 25 remain. This reduction to 11% of the images, will cause further processing, such as localization to be sped up by a factor 9. The number of images taken from inside the rooms, that are left in the CDS,

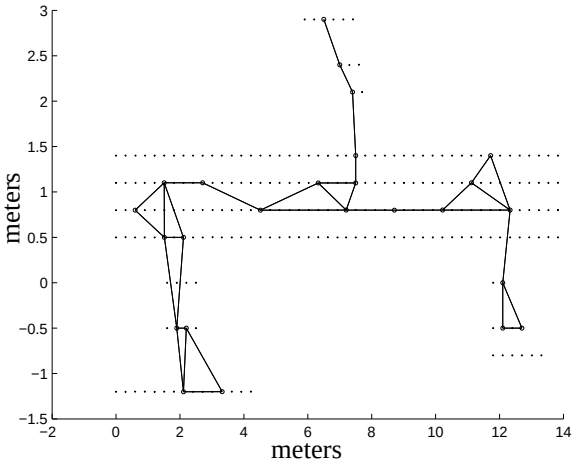

Figure 5: The pruned graph obtained by applying the CDS-algorithm on the complete correlation graph depicted in figure 4.

is relatively higher than the number of images from the corridor, indicating that the image set correctly represents the environment.

6.2 Localizing new images in the topological map

To test the quality of the computed set of images we will use it for a localization task. To localize itself in the topological map the robot will take a new image and perform the same matching technique as used for the calculation of the correlation between the images, described in Section 5. Hence, we can define the existence of an edge between two nodes in the original graph, as the ability to match one of the nodes location given that the other node is in the remaining set. It should be obvious that all the original image-locations can be localized in the topological map again, because the property of the CDS insists that every node is in the remaining set or can be reached by one edge from a node in the remaining set.

Of course a better way to evaluate the CDS algorithm, is to test it on a new image. We will simulate this with the leave-one-out method in the following manner. First we will leave one node with its edges out of the original graph, which will be regarded as a new image that needs to be localized in the graph. Then a CDS will be computed from the remaining nodes. We can now test if the new image can be matched with a remaining image by checking if there is an edge in the original graph to one of the nodes in the computed CDS.

We repeated this scheme for all the images in the data set and found that 226 of the 234 (97%) could be localized in the appropriate graph, see

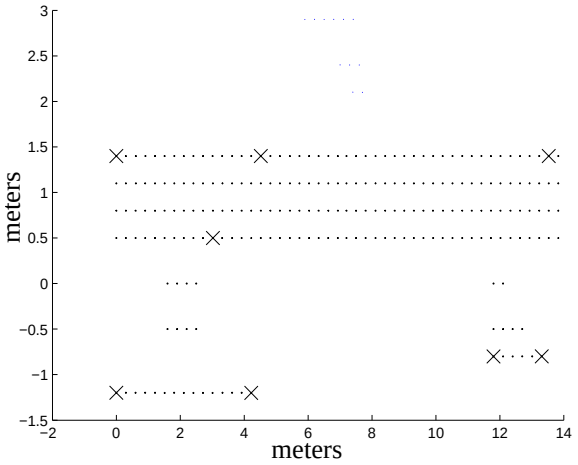

Figure 6: Result of the test to match new images with a pruned topological map. An X indicates an image that could not be matched with the CDS. All other images denoted by the dots can be correlated.

figure 6. Also, most of the images that could not be localized in the topological map were taken at the geometrical border of the robots domain. Of course a robot should not be expected to be able to localize itself outside the internal model.

7 Conclusion

In this paper we described the application of a Connected Dominating Set approximation on the task of appearance based robot localization. Using this technique the performance of on-line localization can be improved significantly, by reducing the number of images in the dataset, while trying to preserve the total information in the set. We have successfully applied the method on a set of panoramic images taken by our robot. It could however also be used to enhance other appearance based applications, such as object-recognition systems. However it has yet to be shown that the method can also be applied to a set of ordinary (not panoramic) images. The resulting appearance based graph will be a little awkward because there is no one-to-one mapping between the images and the locations in the environment. One way of tackling this is by assigning a node for each set of images taken at one position, thus using a little positional information.

The resulting pruned graph (see figure 5) seems to be also useful for efficient path-planning, since all rooms are still reachable while at the same time the total number of paths is reduced substantially. In general however this is not a good idea, considering that the CDS algorithm minimizes the number of images in such a way that every location should still be reachable by one path. Consequently, the pruned graph is likely not to represent large loops present in the environment, because in a loop nodes are reachable by two paths. This will make subsequent path planning suboptimal.

The proposed algorithm is meant for off-line use, that is: the complete set of images of the whole environment is provided to the algorithm at once. It would be useful to also have an on-line algorithm for rejecting images, in which the image dataset is composed incrementally by an exploring robot. This is considered to be future work.

Acknowledgment

The work described in this paper was conducted within the EU Integrated Project COGNIRON ("The Cognitive Companion") and was funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020.

References

[1] B. Bakker, Z. Zivkovic, and B. Krose. Hierarchical dynamic programming for robot path planning. *Technical Report, University of Amsterdam (Submitted to IROS 2005)*, 2005.

[2] R. Basri, E. Rivlin, and I. Shimshoni. Visual homing: Surfing on the epipoles. *International Journal of Computer Vision*, 33(2):117–137, 1999.

[3] R. Bunschoten. *Mapping and Localization from a Panoramic Vision Sensor.* PhD thesis, University of Amsterdam, November 2003.

[4] A.J. Davison and D.W. Murray. Mobile robot localization using active vision. *In Proc. 5th European Conference on Computer Vision, Germany*, 1998.

[5] G. Dudek and D. Jugessur. Robust place recognition using local appearance based methods. In *IEEE Conf. on Robotics and Automation*, pages 1030–1035, San Francisco, April 2000.

[6] S. Guha and S. Khuller. Approximation algorithms for connected dominating sets. In *ESA '96: Proceedings of the Fourth Annual European Symposium on Algorithms*, pages 179–193. Springer-Verlag, 1996.

[7] L. J. Guibas, R. Motwani, and P. Raghavan. The robot localization problem. In Goldberg,

Halperin, Latombe, and Wilson, editors, *Algorithmic Foundations of Robotics, The 1994 Workshop on the Algorithmic Foundations of Robotics, A. K. Peters*, 1995.

[8] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision, secon edition.* Cambridge University Press, 2003.

[9] M. Jogan and A. Leonardis. Robust localization using an omnidirectional appearance-based subspace model of environment. *Robotics and Autonomous Systems, Elsevier Science*, 45(1):51—72, 2003.

[10] B. J. A. Kröse, N. A. Vlassis, R. Bunschoten, and Y. Motomura. A probabilistic model for appearance-based robot localization. *Image Vision Comput.*, 19(6):381–391, April 2001.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[12] E. Remolina and B. Kuipers. Towards a general theory of topological maps. *Artificial Intelligence*, 152(1):47–104, 2004.

[13] P. L. Sala, R. Sim, A. Shokoufandeh, and S. J. Dickinson. Landmark selection for vision-based navigation. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 3131–3138, Sendai, Japan, September 2004. IEEE/RSJ, IEEE Press.

[14] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, volume 1, pages 414–431. Springer-Verlag, 2002.

[15] S. Se, D.G. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research*, 8(21):735–758, 2002.

[16] R. Sim and G. Dudek. Learning and evaluating visual features for pose estimation. *In Proc. International Conference Computer Vision*, 1999.

[17] S. Thrun and A. Buecken. Learning maps for indoor mobile robot navigation. Technical Report CMU-CS-96-121, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, 1996.

[18] P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *IJCV*, 24(3):271–300, 1997.

[19] I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of ICRA 2000*, volume 2, pages 1023 – 1029, April 2000.

[20] Z. Zivkovic, B. Bakker, and B. Krose. Hierarchical map building using visual landmarks and geometric constraints. *Technical Report, University of Amsterdam (Submitted to IROS 2005)*, 2005.