A geometrically constrained image similarity measure for visual mapping, localization and navigation

Ben Kröse   Olaf Booij   Zoran Zivkovic

Intelligent Systems Laboratory, University of Amsterdam, Amsterdam, the Netherlands

`(krose,obooij,zivkovic)@science.uva.nl`

Abstract— **This paper presents an measure for image similarity based on local feature descriptions and geometric constraints. We show that on the basis of this similarity an appearance graph representation of the environment of a mobile robot can be made. This graph can be used for representing semantic information about the space, and can be used for visual navigation. The image similarity measure is robust for occlusions by people in the neighbourhood of the robot.**

Index Terms— **Visual mapbuilding, localization, navigation**

I. INTRODUCTION

An internal representation of the environment is needed for optimal mobile robot navigation. Traditionally such a model is represented as a geometric model indicating admissible and non-admissible areas. The robot has to know its location within such a model and in most of the times has to estimate the parameters of the models simultaneously (SLAM).

Now cameras and processing power are becoming cheaper, visual information is used more often in environment modeling. For example, visual features are used to solve the loop closing problem in geometric SLAM. A step further are approaches which model the environment only in appearances, in contrast to explicit geometric representations of space.

In this paper we present our recent work on appearance modeling of the environment. On the basis of a set of omnidirectional camera images an 'appearance graph' is constructed. This graph can be used for navigation and for a categorization. A prerequisite for making the graph is a good similarity measure between images. The paper first present a brief overview on visual perception and space models. Then the work on appearance modeling in robotics is summarized. Section IV presents the graph based model. The similarity measure and the applications of the graph are presented in sections V,VI and VII. The robustness for visual occlusions is presented in section VIII.

II. SPATIAL REPRESENTATIONS AND VISUAL INFORMATION

Work on spatial representations has been carried out in various fields. From the field of behavioural psychology, the early studies of Tolman [24] using rats in various mazes, showed that rats could learn a 'cognitive map' and reason with that representation. Also from the field of neuroscience a cognitive map theory was presented by O'Keefe and Nadel in 1978 [2]. The theory focussed on hippocampal functioning and suggests
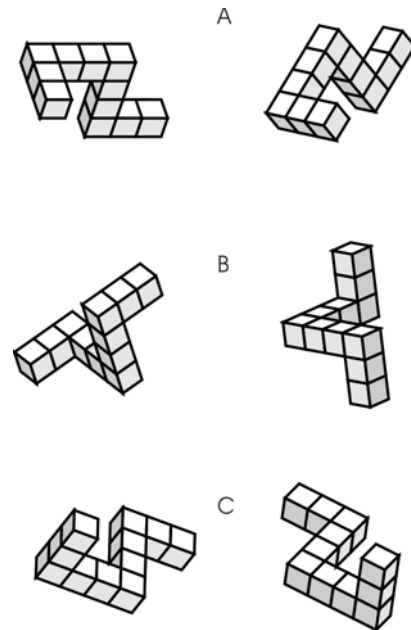

Fig. 1. Shepard & Metzlers Mental rotation task. Subjects were shown pairs of drawings of three-dimensional objects and asked whether the members of a pair were identical. The task can be solved for physical objects by rotating one of them until they can be viewed from the same perspective, but in this case the subjects had to perform the rotation "mentally".

that this brain structure is the core of an extensive neural system subserving the representation and use of information about the spatial environment. The authors describe that visual cues play an important role in map learning. An intriguing debate took place the end of the 1980's, when Kosslyn [9] presented his theory on mental imagery. In his research he studied to which extend images serve as data structures for human memory. As a part of that work spatial representations were considered. Experiments carried out earlier by Shephard (see figure 1) showed that a in order to judge whether two observations were coming from a same object, the subjects 'mentally rotated' one of shapes and compared it to the other: the matching was done in the image domain instead of in the 3D shape domain .

Also in other fields, for example engineering, studies have been carried out on the representation of space. From the field of city design, 'cognitive maps' describe hoe people perceive and understand the environment [13]. Lynch's stud-

ies were based on 'salient' objects, or buildings which are (visually) perceived by humans. Also in the field of machine intelligence 'cognitive maps' have been introduced. Kuipers [11] defines a cognitive map as a layered model consisting of the identification and recognition of landmarks and places from local sensory information, control knowledge of routes, a topological model of connectivity and metrical descriptions of shape, distance, direction, orientation, and local and global coordinate systems.

In the literature on spatial representations in humans, there is clearly evidence that visual information might be a basis of spatial models.

III. APPEARANCE BASED REPRESENTATIONS FOR ROBOTICS

Traditionally robots use a two-dimensional, geometrically accurate representation of a three-dimensional space; a 'mapping' from world coordinates to an indicator which tells whether the position is occupied or not. Sensors to make such maps are typically range sensors such as sonar or laser range scanners. Scanning range sensors make it possible to make 3D geometric representations, sometimes augmented with appearance information from a camera. More recently computer vision techniques are presented. Sets of images (structure from motion) are used to make 3D representations. Dense methods have been presented [5], as well as 3D reconstruction from local salient features ('landmarks') [20],[17]. On-line simultaneous localization and reconstruction of visual landmark positions was presented in [1] but currently only for small scale environments.

In addition to the metric mapping it is common to represent the environment in terms of a topological map: distinct places are coded as nodes in a graph structure with edges which encode how to navigate from one node to the other [6]. The nodes and the edges are usually enriched with some local metric information. Mostly such topological maps are derived from geometric maps. However, recently also visual information has been used to characterize nodes [8], [23].

All approaches described above use vision to reconstruct a 3D (or 2D) representation of the environment of the robot. The question is, whether we can also use models that do *not* try to recover a 2D or 3D representation of space?

In machine vision, *appearance modeling* of objects was introduced about 10 years ago [14]. Nayar showed that an object could be modeled as the set of views of all different poses w.r.t. the camera. In feature space these views form a low dimensional manifold. An unknown object is classified by finding the nearest manifold. Class label and pose are recovered simultaneously (see figure 2). For environment modeling appearance models of space were presented [10]. In these approaches, the environment is modeled as an 'appearance map' that consists of a collection of camera (or other sensor) readings obtained at known poses (positions and orientations). These methods have shown to be able to localize a mobile robot but have the problem that supervised data, consisting of images and corresponding poses, are needed.
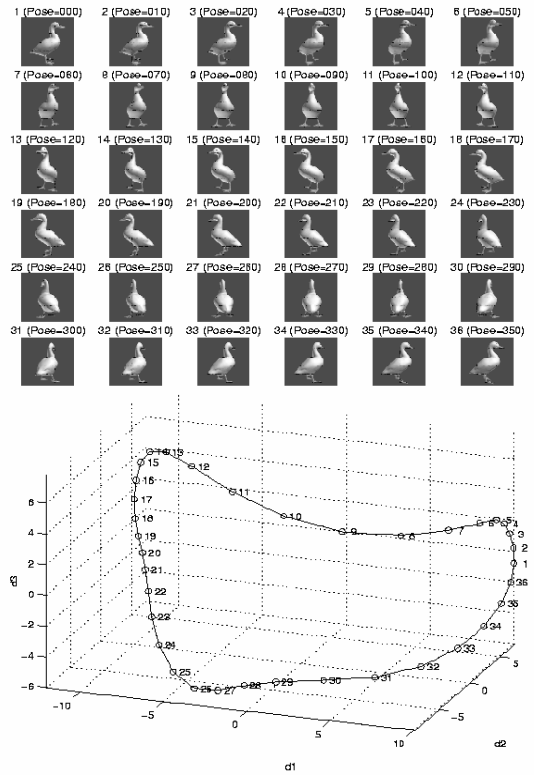

Fig. 2. Appearance modeling of objects. Instead of representing the object as a 3D shape, the object is models as a set of feature vectors derived from views at different poses. These views can be modelled as a curve in a low dimensional feature space. After Murase and Nayar[14].

IV. THE APPEARANCE GRAPH REPRESENTATION

In our current approach on appearance modeling we avoid the problem of a supervised training set. We collect a set of camera images and use this image set to construct an 'appearance graph'.

In an appearance graph each vertex or 'node' represents a pose (which we do not know) and is characterized by the camera image taken at that location. An edge between two nodes is defined if the two images are sufficiently similar. As we will see in the next section, the similarity checks whether it is possible to perform 3D reconstruction of the local space from the two corresponding images. The idea behind this is that we want to have a similarity measure which states that similar images are taken at adjacent positions. The appearance graph contains 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 that are from another space, for example a corridor, that is connected with the room via a narrow passage, for example a door. As the result from n images we obtain 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 similarity of the nodes.

An example of such graph that we obtained from a real data set is given in figure 3. An edge is drawn between

Fig. 3.   The appearance graph after the robot was driven manually along a trajectory starting at A and finishing at A. The nodes indicate locations where omnidirectional images were taken. In order to plot these locations we used odometry readings corrected after the loop closing

two nodes if the similarity exceeds some threshold. Note that in constructing the graph no information about the positions of the nodes is used. In the figure we used these only for visualization.

As can be seen from the graph, rooms are characterized by highly connected parts of the graph. In section VI we describe how to extract such groups of images automatically from the graph (V, S). 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 [3]). Section VII describes how we use the graph for navigation. However, in the next section we will first present the similarity measure.

V. OUR IMAGE SIMILARITY MEASURE

Extensive work on image similarity has been presented in the field of image database retrieval. Methods developed in this field are generally based on local features, or visual *landmarks*. Popular methods try to find a vocabulary of 'codebook' vectors which are used for image matching. Also in the field of robotics these approaches are introduced [15]. Note that these methods remove all geometric information from the similarity measure. Our image similarity measure is based on knowledge that images are taken from a moving camera in an environment. This is exactly what robotics makes different from most image retrieval methods.

As mentioned earlier, our image similarity measure denotes that a 3D reconstruction is possible between the images. The method for 3D reconstruction is based on local salient features as landmarks. Currently we use the SIFT feature detector [12]. The SIFT feature detector extracts 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 robust to some light changes, which makes it

appropriate for our application. The method for computing the similarity between two images is split in two parts:

1) Are there matching landmarks in the two images, and
2) Do these landmarks fulfill the epipolar constraint?

A. Matching Landmarks

Visual landmarks are used often in robotics for navigation [20],[17],[16]. It has been shown that it is possible to reconstruct both the camera poses 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 [1] 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 the case described in [18], [19]. First we check if there are many similar (repetitive) landmarks within each image separately. Such landmarks could potentially lead to false matches. We discard those landmarks that have 6 or more similar instances in the same image.

Then, 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 Euclidian distance between the landmark descriptors. If the goodness of the second best match is less then 0.8 of the best one it means that the match is very distinctive. According to the experiments in [12] this typically discards 95% of the false matches and less then 5% of the good ones. This is repeated for each pair of images and it is very computationally expensive. Fast approximate methods were discussed in [12].

B. Geometric Constraints

After finding the possible matches for each pair of images from our data set as described above, we apply a geometric constraint. Let there be N matching landmark points between images m and l. The image positions of the points in the m-th image in homogenous coordinates are denoted as $\{\mathbf{x}_m^{(1)}, ..., \mathbf{x}_m^{(N)}\}$. The corresponding points in the l-th image are $\{\mathbf{x}_l^{(1)}, ..., \mathbf{x}_l^{(N)}\}$. If the i-th point belongs to the static scene, then, for a projective camera, the positions are related by:

$$(\mathbf{x}_m^{(i)})^T E \mathbf{x}_l^{(i)} = 0 \qquad \text{for all } i. \tag{1}$$

where the matrix E is also known as the 'essential matrix'. Estimating E is an initial step for 3D reconstruction of the space from images; here we use it for computing image similarity.

The approach described above for feature matching may lead to initial false matches. Standard robust M-estimators can deal with a certain amount of outliers. The robust algorithm called RANSAC is usually used [7] if there are more outliers. It was shown [25] that a combination that performs the best is when the RANSAC is used first and then the M-estimator. Instead of following the [7] completely we use only the distinctive matches as in [12] 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

Fig. 5. The clustering found with the graph cut mechanism reflects the structure in the building
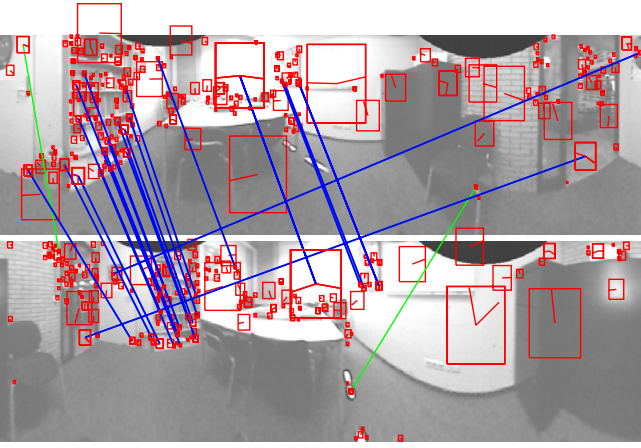
Fig. 4. Matching two images. The red boxes indicate the SIFT features found in the images. The lines connecting two of these features indicate that they correspond. If the line is blue, this means the corresponding pair agrees with the epipolar constraint. If it is green, it does not agree with the epipolar constraint and is thus probably an outlier.

algorithm [7] which requires at least 8 matching points. With such small number of false matches it is possible to use the robust M-estimator directly. The whole procedure goes as follows:

- extract SIFT landmarks from all images
- discard self similar landmarks within each image
- find distinctive matches between pairs of images
- if there are more than 8 matches:
- estimate the fundamental matrix using M estimator and RANSAC
- if there are still more than 8 matches then there is an edge in the graph

In [4] we use a ground floor constraint, which means that an edge can even be defined on the basis of 3 matches. An example of the matches in a realistic situation is depicted in figure 4.

VI. CLUSTERING IN THE GRAPH

Using the above presented similarity measure, we are able to make a graph-like representation of the environment. The graph can be considered as a low-level topological map, with the vertices (nodes) indicating omnidirectional images and the links indicating a similarity between the images. By clustering in this representation we are able to come to a higher level topological map, with clusters indicating regions in which the nodes are very similar.

The graph (V, S) gives this information about the structure of the environment. Convex spaces contain nodes which are highly interconnected, and doorways will have nodes with fewer connections. The graph V can be divided into subsets by cutting a number of edges. There exist different graph cut mechanisms. In [26] we present our approach which is a fast approximate solution to the normalized graph cut method from [21]. In figure 5 it can be seen that the method results in meaningful clusters. The nodes inside the rooms get the same label, and each room gets its own label. The hallway is divided

into four regions, which indicates that the appearance is not uniform in the hallway.

We use the clustered representation to obtain a semantic description of space by Human Robot Interaction. In [22] we describe a situation where the robot is guided around by a user, while the user occasionally gives a label to a location (for example: 'corridor', 'living room'). The image taken at that location is labeled with that label. By using our clustering method, all images (nodes) in a cluster obtain the same label.

VII. NAVIGATION USING THE GRAPH

The appearance graph can also be used for navigation. The challenge is that the graph does not contain any metrical information: only appearance information and neigbourhood relations. In [4] we present our navigation method. It is based on two steps. First we define a *cost function* on the appearance graph, indicating for each node in the graph the distance to the goal node, and then we use a greedy visual navigation mechanism to drive to the goal.

A. Cost function

We assume that the goal location is given by a node in the graph. This can be for example be done by the user giving a semantic label ('kitchen') which corresponds to a node. First Dijkstra's shortest path algorithm is used to compute the cost function, which is the distance D_i from every node i in the graph to this goal node. This algorithm requires the links of the graph to be labeled with a distance measure while we have a similarity measure. Therefore we define the distance d as $d_{ij} = \frac{1}{S_{ij}}$. The distances of the nodes to goal node are used during driving as a heuristic to drive in the direction of the goal node. An example of such a cost function is depicted in figure 6.

B. Greedy visual navigation

At the start of the trajectory the robot localizes itself in the appearance based graph by taking a new observation and comparing it with all the images in the graph following the same matching procedure as used for constructing the graph . The node of the graph with the highest similarity is chosen as the current subgoal node c of the robot. This procedure

Fig. 6. A cost function is created on the basis of distances to the goal node

is linear in the number of nodes and could thus be time consuming.

If a subgoal node is determined the robot tries to pick a new subgoal by comparing the newest observation with all the neighbors of node c that have a smaller distance D_c to the goal node. If one of these images matches, it becomes the new current subgoal c. This procedure is repeated for the neighbors of the new c, until the node is found that is closest to the goal node and does still robustly match the new observation (see figure 7

When a subgoal is determined, the heading is estimated in order to drive in its direction. This heading will not be perfectly directed toward the subgoal, partly because of sensor-noise, but also because the environment could have changed after the appearance based map was constructed. Therefore a recency weighted averaging filter is used which to takes into account previous estimates of ϕ.
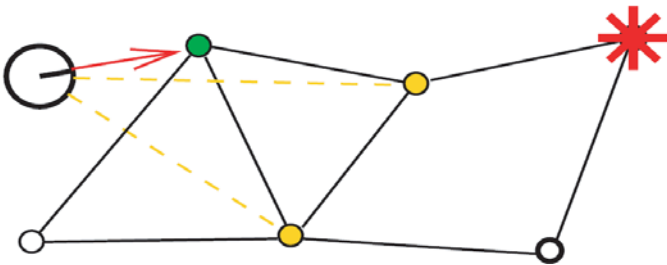
Fig. 7. From the current location the robots heads towards an intermediate target (green). In parallel the similarity with neigbour nodes closer to the goal are computed (yellow). If these nodes can be reached from the current position, the robot will head to a new intermediate target

C. Heading estimation using the epipolar geometry

The navigation method requires that the robot is able to compute a heading direction on the basis of two images. To compute this heading direction we use the same mechanism described in section V. On the basis of the current image 1 and the target image 2 we can again compute the essential matrix:

$$(\mathbf{x}_1^{(i)})^T E \mathbf{x}_2^{(i)} = 0 \qquad \text{for all } i. \tag{2}$$

TABLE I
PATH LENGTHS IN METERS WITH PEOPLE BLOCKING THE VIEW

#Persons	path lengths
0	14.2
1	15.2
2	18.0
3	19.0
4	23.6

The essential matrix bears the relative rotation R and translation $\vec{t}$ up to an unknown scale between the positions of the two images as follows:

$$E = RS, \tag{3}$$

where S is a skew-symmetric matrix composed of the elements of $\vec{t}$. The essential matrix can be decomposed into 4 different solutions of $\vec{t}$ and R, and we use standard methods to select the correct one. The heading ϕ the robot has to drive when navigating from the current image to the target image can be calculated using

$$\phi = \text{atan2}(t_y, t_x). \tag{4}$$

A movie showing the visual navigation can be obtained from the authors[1].

VIII. EXPERIMENTS ON VISUAL OCCLUSIONS

Many localization methods fail if the environment has changes due to new furniture of occlusions by surrounding people. To put more strain on the visual navigation method we now test the ability to drive while part of the view is blocked by people walking next to and in front of the robot. See Figure 8 for an indication of the view the robot has while 4 persons are standing next to it. The persons are walking very near the robot at more or less 20 cm distance. The path that had to be traversed is the same as one of the paths in the previous section. Tests are conducted with respectively 1, 2, 3 and 4 persons.

The robot still managed to reach the goal location in all 4 tests. Nonetheless it was clear that for every person that was added, the navigation was a bit more difficult. In table I it is shown that the path length increases if a larger part of the view is blocked. This is not only caused by small divergences of the correct path, but also because the robot sometimes took a longer route around the pillar in the hallway. Surprisingly the robot never had to use its recovery method. The number of times that the heading to the subgoal could not be estimated did increase though. For one and two persons it could still match 100% of the observations, but this decreased to 90% for the runs with three person and four persons.

During the test with 4 persons an additional thing happened. Because no collision avoidance was used and the robot was sometimes heading for a doorpost or the pillar, we had to stop it manually and push it back. This happened 3 times.

[1]Currently the video is available from http://staff.science.uva.nl/~krose/movies/

Fig. 8. Four persons blocking the view of the robot.

IX. Conclusions

We presented a system for localization, mapping and navigation on the basis of appearance data from an omnidirectional vision system. The robot is able to find and traverse paths in the visual domain and can navigate from one state to the other. Navigation proved to be robust in a dynamic environment with people walking close to the robot.

The experiments we presented in this paper showed that the robot can successfully map and navigate in a region of about 500 m^2. We realize that this is relatively small compared to current work on SLAM methods, in which paths of more than some kilometers are traveled. On the other hand, the application domain we work on is a personal robotic assistant in a domestic environment, where robustness to dynamics in the environment, light conditions and interactions with people are far more important.

Currently we are integrating the navigation approach in a more complete robot system that incorporates people detection, people following and exploration. All these methods make use of the same omnidirectional vision system.

Acknowledgements

References

[1] A.J.Davison and D.W.Murray. Mobile robot localization using active vision. *In Proc. 5th European Conference on Computer Vision, Germany*, 1998.
[2] J. O'Keefe L. Nadel. *The Hippocampus as a Cognitive Map*. Oxford University Press, 1978.
[3] R. Basri, E. Rivlin, and I. Shimshoni. Visual homing: Surfing on the epipoles. 1998.
[4] O. Booij, B. Terwijn, Z. Zivkovic, and B. Kröse. Navigation using an appearance based topological map. In *Proc. IEEE International Conference on Robotics and Automation*, pages 3927–3932. IEEE, 2007.
[5] R. Bunschoten and B. Kröse. 3-D scene reconstruction from cylindrical panoramic images. *Robotics and Autonomous Systems (special issue)*, 41(2/3):111–118, November 2002.
[6] E.Remolina and B.Kuipers. Towards a general theory of topological maps. *Artificial Intelligence*, (152):47–104, 2004.
[7] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision, secon edition*. Cambridge University Press, 2003.
[8] Ulrich I. and Nourbakhsh I. Appearance-based place recognition for topological localization. *IEEE Int. Conf. on Robotics and Automation*, pages 1023–1029, 2000.
[9] S. Kosslyn. *Image and Mind*. Harvard University Press, Cambridge, MA, 1980.
[10] B.J.A. Kröse and R. Bunschoten. Probabilistic localization by appearance models and active vision. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2255–2260, 1999.
[11] B. J. Kuipers. Modeling spatial knowledge. *Cognitive Science*, 2:129–153, 1978.
[12] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91–110, 2004.
[13] K. Lynch. *The Image of the City*. Harvard University Press, Cambridge, 1960.
[14] H. Murase and S. K. Nayar. Visual learning and recognition of 3-d objects from appearance. *Int. Jrnl of Computer Vision*, 14:5–24, 1995.
[15] P. Newman, D. Cole, and Kin Ho. Outdoor slam using visual appearance and laser ranging. In *International Conference on Robotics and Automation*, 2006.
[16] R.Sim and G.Dudek. Learning and evaluating visual features for pose estimation. *In Proc. International Conference Computer Vision*, 1999.
[17] P. Sala, R. Sim, A. Shokoufandeh, and S. Dickinson. Landmark selection for vision-based navigation. *In Proc. International Conference on Intelligent Robots and Systems, Japan*, 2004.
[18] 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.
[19] F. Schaffalitzky and A. Zisserman. Automated location matching in movies. *Computer Vision and Image Understanding*, 92:236–264, 2003.
[20] 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.
[21] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
[22] Thorsten Spexard, Shuyin Li, Britta Wrede, Jannik Fritsch, Gerhard Sagerer, Olaf Booij, Zoran Zivkovic, Bas Terwijn, and Ben Kröse. Biron, where are you? - enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 934–940. IEEE, October 2006.
[23] A. Tapus, S. Vasudevan, and R. Siegwart. Towards a multilevel cognitive probabilistic representation of space. *In Proc. of the International Conference on Human Vision and Electronic Imaging X, part of the IST-SPIE Symposium on Electronic Imaging*, 2005.
[24] E.C. Tolman. Cognitive maps in rats and men. *The Psychological Review*, 55(4):189–208, 1948.
[25] P. Torr and D. Murray. The development and comparison of robust methods for estimating the fundamental matrix, 1997.
[26] Z. Zivkovic, B. Bakker, and B.J.A. Kröse. Hierarchical map building using visual landmarks and geometric constraints. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 7–12, 2005.