Vision-Based Localization for Mobile Platforms

Josep M. Porta and Ben J.A. Kröse

IAS Group, Informatics Institute, University of Amsterdam,
Kruislaan 403, 1098SJ, Amsterdam, The Netherlands
{porta,krose}@science.uva.nl

Abstract. In this paper, we describe methods to localize a mobile robot in an indoor environment from visual information. An *appearance*-based approach is adopted in which the environment is represented by a large set of images from which features are extracted. We extended the appearance based approach with an *active vision* component, which fits well in our probabilistic framework. We also describe another extension, in which depth information is used next to intensity information. The results of our experiments show that a localization accuracy of less then 50 cm can be achieved even when there are un-modeled changes in the environment or in the lighting conditions.

1 Introduction

Localization and tracking of moving objects is one of the central issues in research in intelligent environments. In many cases, persons need to be localized for security or services. Also the localization of movable intelligent embedded systems is an important issue in wireless local networks or for localizing Personal Digital Assistants (PDA's) which can serve, for instance, as museum guides. In this paper, we focus on the localization of a robot platform (the domestic service robot 'Lino' [1], [4] (see Figure 1), developed within the 'Ambience' project), but the described techniques are applicable in many other domains.

Different solutions have been presented for localizing mobile objects. One class of solutions is to use sensors placed in the environment. Such sensors may be cameras, infra-red detectors or ultrasonic sensors. The main problem here is the *identity uncertainty*: the cameras (or other systems) have to infer the identity of the object/person to be localized from the sensor readings and this is, in general, difficult. For instance, the camera-based localization system described in [6] fails when tracking more than three persons. A second class of solutions is to use radio-frequency signals. These signals are naturally present in mobile computing networks, and their identification is trivial. However, accurate localization is difficult due to reflections, absorption and scattering of the radio waves. A localization accuracy of approximately 1.5 meters is reported using this approach combined with a Hidden Markov Model [7].

Localization without any active beacons or without any environment sensor traditionally takes place in the research area of autonomous robots [2], [5].

Fig. 1. The user-interface robot Lino.

Generally, robot localization needs some sort of internal model of the environment based on sensor readings from which the pose (position and orientation) of the robot is estimated. Traditionally range sensors (ultrasonic or laser scanners) were used, but recently much progress has been achieved in localization from vision systems. Although the methods show good results, they are not robust to un-modeled changes in the environment and they are quite sensitive to changes in illumination. To cope with these problems, in this paper we describe two modifications to our localization system: an *active* vision method and the use of depth information for localization These two improvements are embedded in a probabilistic framework. In the following section, we will briefly describe our way of modeling the world and, next, our probabilistic localization framework. Then, we introduce our active vision method and the use of depth information for localization. After this, we described the experiments we performed to validate our contributions and the results we obtained from them.

2 Appearance-Based Environment Representation

In the literature on mobile robots, methods for environment representation come in two flavors: *explicit* or *implicit*. The explicit representations are based on geometric maps of free space sometimes augmented with texture information, i.e., CAD models, or maps with locations of distinct observable objects called landmarks. This approach relies on the assumption that geometric information such as the position of obstacles/landmarks can be extracted from the raw sensors readings. However, the transformation from sensor readings to geometric information is, in general, complex and prone to errors, increasing the difficulty of the localization problem.

As a counterpart, the *implicit* (or *appearance-based*) representation of the environment [8] has attracted lot of attention recently. In this paradigm, the environment is not modeled geometrically but as an 'appearance map' that consists of a collection of sensor readings obtained at known poses. The advantage of this representation is that the pose of the robot can be determined directly comparing the sensor readings obtained at a given moment with those in the appearance-based map.

We use a vision-based appearance representation built with many images from the environment. A problem with images is their high dimensionality, resulting in large storage requirements and high computational demands. To alleviate this problem, Murase and Nayar [8] proposed to compress images, z, to low-dimensional feature vectors, y, using a linear projection

$$y = W\, z. \tag{1}$$

The projection matrix W is obtained by Principal Component Analysis (PCA) of a supervised training set ($T = \{(x_i, z_i)|\ i \in [1, N])\}$) consisting of images z_i obtained at known poses x_i. We keep the subset of eigenvectors that represent most of the variance of the images and we use them as rows of the projection matrix W. After the dimensionality reduction, the map used by the robot for localization $M = \{(x_i, y_i)|\ i \in [1, N])\}$ consists of a set of low-dimensional (typically around 20-D) feature vectors y_i corresponding to the images taken at poses x_i. The use of features instead of raw images saves a large amounts of memory space.

For localization, the robot first has to project the image which is observed at the unknown location to a feature representation. Then, the probabilistic model described next is used to localize the system.

3 A Probabilistic Model for Localization

The camera system is mounted on a pan-tilt device, rigidly attached to the mobile robot. We assume that the orientation of the camera with respect to the robot is given with sufficient accuracy by the pan-tilt device. The absolute pose of the camera is considered as a stochastic (hidden) variable x. The localization method aims at improving the estimation of the pose x_t of the camera at time t taking into account the movements of the robot/head $\{u_1, \ldots, u_t\}$ and the observations of the environment taken by the robot $\{y_1, \ldots, y_t\}$ up to that time[1]. Formally, we want to estimate the posterior $p(x_t|\{u_1, y_1, \ldots, u_t, y_t\})$. The Markov assumption states that this probability can be updated from the previous state probability $p(x_{t-1})$ taking into account only the last executed action, u_t, and the last observation, y_t. Thus we only have to estimate $p(x_t|u_t, y_t)$. Applying Bayes we have that

$$p(x_t|u_t, y_t) \propto p(y_t|x_t)\, p(x_t|u_t), \tag{2}$$

[1] In our notation, the Markov process goes through the following sequence: $x_0 \xrightarrow{u_1} (x_1, y_1) \xrightarrow{u_2} \ldots \xrightarrow{u_t} (x_t, y_t)$.

where the probability $p(x_t|u_t)$ can be computed propagating from $p(x_{t-1}|u_{t-1}, y_{t-1})$

$$p(x_t|u_t) = \int p(x_t|u_t, x_{t-1})\, p(x_{t-1}|u_{t-1}, y_{t-1})\, dx_{t-1}. \tag{3}$$

We discretize equation 3 using an *auxiliary particle filter* [9]. In this approach, the continuous posterior $p(x_{t-1}|u_{t-1}, y_{t-1})$ is approximated by a set of I random samples, called particles, that are positioned at points x^i_{t-1} and have weights π^i_{t-1}. Thus, the posterior is

$$p(x_{t-1}|u_{t-1}, y_{t-1}) = \sum_{i=1}^{I} \pi^i_{t-1}\, \delta(x_{t-1}|x^i_{t-1}), \tag{4}$$

where $\delta(x_{t-1}|x^i_{t-1})$ represents the delta function centered at x^i_{t-1}. Using this, the integration of equation 3 becomes discrete

$$p(x_t|u_t) = \sum_{i=1}^{I} \pi^i_{t-1}\, p(x_t|u_t, x^i_{t-1}), \tag{5}$$

and equation 2 reads to

$$p(x_t|u_t, y_t) \propto p(y_t|x_t) \sum_{i=1}^{I} \pi^i_{t-1}\, p(x_t|u_t, x^i_{t-1}). \tag{6}$$

The central issue in the particle filter approach is how to obtain a set of particles (that is, a new set of states x^i_t and weights π^i_t) to approximate $p(x_t|u_t, y_t)$ from the set of particles x^i_{t-1}, $i \in [1, I]$ approximating $p(x_{t-1}|u_{t-1}, y_{t-1})$. In [9] and [12], you can find details on how this is achieved in the approach we use.

The probability $p(x_t|u_t, x_{t-1})$ for any couple of states and for any action is called the *action model* and it is inferred from odometry. On the other hand, $p(y_t|x_t)$ for any observation and state is the *sensor model*. This probability can be approximated using a nearest-neighbor model that takes into account the J points in the appearance-based map that are more similar to the current observation (see [12]).

4 Active Localization

Traditionally, appearance-based localization has problems in dynamic environments: modifications in the environment are not included in the model and can make recognition of the robot's pose from the obtained images very difficult. To alleviate this problem, we introduce the use of an active vision strategy. Modifications in the environment would only be relevant if the camera is pointing towards them. If this is the case, we can rotate the cameras to get features in other orientations hopefully not affected by the environment changes. Therefore, in case of observations that do not match with those in the appearance-based

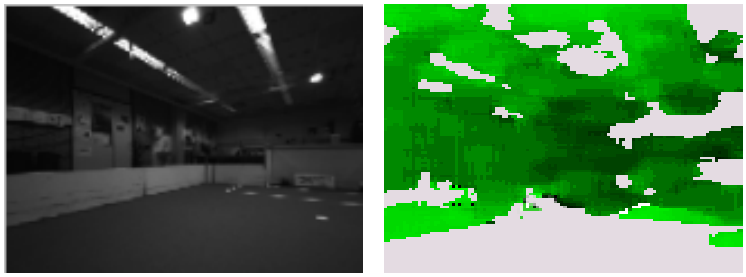

Fig. 2. Plain image (left) and the corresponding disparity map (right). In the disparity map, light gray areas are missing values.

map, the location of the robot can be found out efficiently by issuing the adequate sequence of camera rotations. The question here is how to determine this *adequate* sequence of camera movements.

In [10], we present a method based on the minimization of the estimated entropy $H(u)$ of the stochastic variable x if action u was executed. We describe how to approximate the entropy $H(u)$ by taking advantage of the two main components of our localization system: the particle filter (to get a discrete set of the possible placements of the robot after executing action u) and of the appearance-based map (to get a discrete set of possible observations after executing action u). These two approximations allow to discretize the computation of the entropy $H(u)$, making it feasible.

5 Adding Depth Information

Vision-based appearance localization is sensitive to illumination conditions. An adequate preprocessing of the images such histogram equalization makes the system more robust, but does not solve all problems. Therefore, we decided to complement the visual information with depth information.

For that, we used a commercially available stereo system [3], that provides information of the distance to the nearest object for each pixel in the form of a disparity value obtained matching pixels from the two stereo images. The algorithm we use applies many filters in this matching process both to speed it up and to ensure the quality of the results. For instance, if the area around a given pixel is not textured enough it would be very difficult to find a single corresponding point in the other image: we are likely to end up with many pixels with almost the same probability of being the corresponding point to the pixel we are trying to match. For this reason, pixels on low textured areas are not even considered in the matching process. The result of this and other filtering processes is to produce a sparse disparity map: a disparity map where many pixels don't have a disparity value (see Figure 2). This makes the use of standard PCA to determine the projection matrix unfeasible and we have to use more elaborated techniques such as the EM algorithm we introduced in [11].

Once we have a way to define features from disparity maps, it remains the question of how to combine the information coming from disparity with that obtained from intensity to defined a unified sensor model. Two possible solutions come to mind: to combine them in a conjunctive way or in a disjunctive one.

A conjunctive-like combination can be achieved factorizing the sensor model

$$p(y_d, y_i|x) = p(y_d|x)\, p(y_i|x), \tag{7}$$

with y_d the features obtained for disparity and y_i those for intensity. In this way, only those training points consistent with both the current intensity image and the current disparity map are taken into account to update the robot's position. The problem of this formulation is that wrong matches for intensity or for disparity would result in an almost null sensor model and, thus, the position of the robot would be updated almost without sensory information.

To avoid this, we propose to use a disjunctive-like model that can be implemented defining the global sensor model as linear combination of the intensity and disparity sensors models

$$p(y_d, y_i|x) = w_d \sum_{j=1}^{J} \lambda_j\, \phi(x|x_j) + w_i \sum_{j=1}^{J'} \lambda'_j\, \phi(x|x'_j), \tag{8}$$

where x_j and x'_j are the training points with features more similar to those of the current disparity and intensity image respectively. The weights w_d and w_i can be used to balance the importance of the information obtained with each type of sensor. If both information sources are assumed to be equally reliable, we can set $w_d = w_i = 1/2$.

With this expression for the sensor model, all the hypotheses on the robot position suggested by the current observation are taken into account. The particle filter [9] [12] we use to update the probability on the robot's position takes care of filtering the noise and, thus, of preserving the hypothesis that is more consistent over time.

6 Experiments and Results

In this section, we describe the experiments we performed to validate the our contributions both on active localization and on localization using disparity maps.

6.1 Experiments on Active Localization

We tested our action evaluation system in an office environment. We mapped an area of 800 × 250 cm taken images every 75 cm and every 15 degrees. This makes a total amount of about 400 training images. The short distance between training points make images taken at close positions/orientations to look very similar. In the experiments, we compress the images using PCA keeping 5 feature detectors, we use 10 nearest neighbors to approximate the sensor model $p(y|x)$,

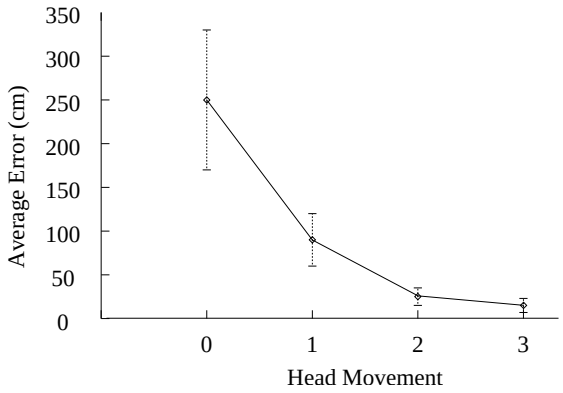

Fig. 3. Evolution of the average error (and the standard deviation) w.r.t. the correct position as we get new images.

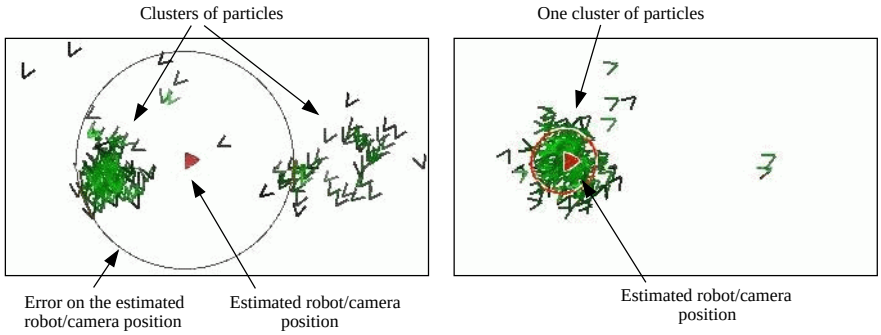

Fig. 4. Convergence toward the correct position estimation: particles around different hypotheses (left) and particles around a single correct hypothesis (right).

and we define the initial distribution $p(x_0)$ uniformly over the configuration space of the robot. We considered 22 different orientations for the camera and we used up to 150 particles to approximate $p(x_t|u_t, y_t)$.

We tested the system placing the robot at positions not included in the training set, rotating the camera as measuring the error and $\|c - a\|$ with c the correct position and a the position estimated by our system.

Figure 3 shows the decrease on the average positioning error as new actions are issued. The results shown correspond to the average and the standard deviation over ten runs placing the camera in two different testing positions. We can see that the entropy-based action selection allows a fast reduction of the localization error as the head is moved and new images are processed. If we consider the estimation a to be correct if the closest training point to a is the same as the closest training point to the correct position c, then the success ratio in localization after 3 camera movements is over 95%.

Figure 4 shows a typical evolution of particles from a distribution around different hypothesis (left) to the convergence around the correct position (right) achieved as new images are processed. In the figure, each '>' symbol represents a particle (the brighter the larger its weight) and the triangle represents the pose of the camera. The circle represents the standard deviation of particles in the XY plane.

6.2 Experiments on Localization using Disparity Maps

To test the invariance of the feature detectors obtained from disparity maps to changes in illumination we acquired an appearance-based map in a area of 900×500 cm. Images where collected every 50 cm (both along X and Y) and every 10 degrees. This makes a total amount of about 4000 images.

Table 1. Ratio of average feature detector variation due to illumination vs. the changes due to a small translations.

Image Process	**Illumination Setups**		
	Bulb Lights	**Natural Light**	**Average**
Plain Images	3.85	4.64	4.24
Hist. Equalization	1.50	1.84	1.67
Gradient Filter	1.11	1.37	1.24
Disparity Map	0.68	0.79	0.73

We analyzed the sensitivity of the feature detectors to two of the different factors that can modify them: translations and changes in illumination. For this, we compute the ratio

$$r(a, b, c) = \frac{\|y_c - y_a\|}{\|y_b - y_a\|}, \tag{9}$$

with y_a the feature detectors of image at pose a with the illumination setup used to collect the appearance map (tube lights), y_b the feature detectors of the image obtained with the same orientation and the same lighting conditions but 50 cm away from a, and y_c the image obtained at pose a but in different illumination conditions. The greater this ratio, the larger the effect of illumination w.r.t. the effect of translations and, thus, the larger the possible error in localization due to illumination changes.

We used two illumination setups for the test: bulb lights and natural light (opening the curtains of the windows placed all along one wall of the lab). These two tests sets provide changes both in the global intensity of the images and in the distribution of light sources in the scene, that is the situation encountered in real applications.

We computed the above ratio for the feature detectors obtained from plain images, from disparity maps and also for images processes with two usual tech-

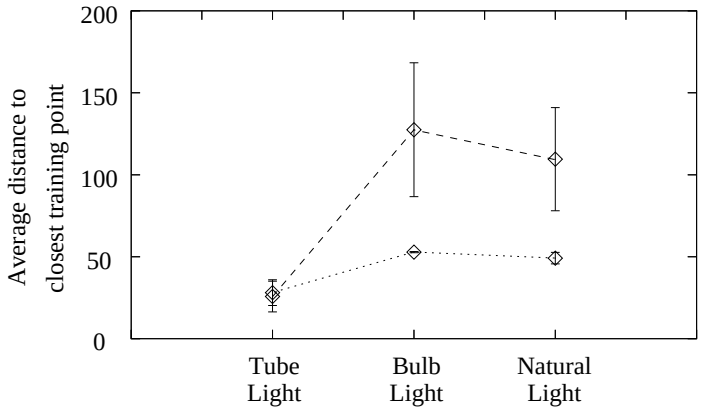

Fig. 5. Error in positioning in three different illumination conditions, using only intensity images (dashed line) and intensity and disparity images (dotted line).

niques for dealing with illumination related problems: histogram equalization and a gradient-based filter.

Table 1 shows the results we obtained for the experiment just described. In this table, we can see that, as expected, using processed images the ratio decreases considerably compared with the ratio using images without any illumination correction. In the case of disparity maps, this ratio is the smallest one meaning that disparity is the best of the three techniques we tested, as far as independence of illumination is concerned.

To assess the contribution of using disparity maps in appearance-based localization, we moved the robot along a pre-defined path in the three different illumination conditions mentioned above: tube lights, bulb lights and natural light. At regular distances along the test path, we took an image and we computed the corresponding sensor model using the J training points with features more similar to those corresponding to the just obtained image. The closer the training points used to define the sensor model to the actual position of the robot, the better the sensor model and, thus, better the update of the robot position estimation.

Figure 5 shows the average and the variance for all the test points all along the path of the error defined as

$$e = \min_{\forall n} \|r - n\|, \tag{10}$$

with $r = (r_x, r_y, r_\phi)$ the poses of the robot at the test position and $n = (n_x, n_y, n_\phi)$ the pose of the points used to define the sensor model, that are different for each test position. An error in the range $[25, 50]$ is quite reasonable since the distance between training points in X and Y dimensions is 50 cm.

We repeat the test in two cases: (a) using only intensity images (dashed line on Figure 5) and (b) using, additionally, disparity maps (dotted line on the

figure). In the first case we use $J = 10$ and in the second case we use $J = 5$ but for both intensity and disparity so, we also get 10 nearest-neighbors.

We can see that the use of disparity maps results in a reduction of the error in the sensor model when illumination is different from that in which the training set was obtained (tube lights). Consequently, the use of feature detectors computed from disparity maps increase the quality of the sensor model and, thus, it helps to obtain a more robust localization system.

7 Conclusions

In this paper, we have introduced two extensions to the traditional appearance-based robot localization framework: active selection of robot actions to improve the localization and used of disparity maps for localization. These two extensions are possible thanks to the use of a stereo camera mounted on the mobile head of the service robot Lino.

The experiments we report with our active vision system show that this mechanism effectively helps to find out the location of the robot. This is of great help in dynamic environments, where existing appearance-based localization system exhibited some problems.

Our second contribution is the use of sparse disparity maps to increase the robustness of appearance-based localization to changes in illumination. The results we have presented show that disparity maps provide feature detectors that are less sensible to changes in the lighting conditions than feature detectors obtained from images processed with other techniques: histogram equalization and gradient-based filters. These techniques work well when we have changes in the global illumination but they do not deal properly with different distributions of light sources. Disparity maps are more consistent over changes in the number and in the position of the light sources because only reliable correspondences are taken into account when defining the disparity map. These reliable matches are likely to be detected in different lighting conditions.

We have shown that, using features from disparity maps in addition to those obtained from intensity images, we can improve the quality of the sensor model when illumination conditions are different from those in which the training set is obtained. Thus, disparity maps are a good option to increase the robustness of appearance-based robot localization.

The good results achieved using disparity maps cames at the cost of using a more complex hardware (we need not only one camera but two calibrated ones) and software (the disparity computation process is more complex than the histogram equalization and the gradient filter processes).

The main assumption behind our approach is the existence of a training set obtained off-line and densely sampled over the space where the robot is expected to move. To obtain this training set is not a problem, but it would be desirable the robot to build it on-line. To achieve this improvement, we have to explore the use of incremental techniques to compress the images obtained as the robot moves in the environment.

The type of environment representation underlying our localization system is computationally very cheap: after the dimensionality reduction the appearance-based map is small and the linear projection of images is an efficient process. For this reason, the localization system could be easily implemented in fields other than autonomous robots as, for instance, PDA's or mobile phones provided with cameras. Since the localization would be performed by the same device to be localized, we avoid the *identity uncertainty* problem, and, additionally, the accuracy in localization that could be achieved is better than that reported using radio-frequency techniques.

Acknowledgments

This work has been partially supported by the European (ITEA) project *"Ambience: Context Aware Environments for Ambient Services"*

We would like to thank Bas Terwijn for helping us to perform the experiments reported on this paper.

References

1. A.J.N. van Breemen, K. Crucq, B.J.A. Kröse, M. Nuttin, J.M. Porta, and E. Demeester A user-interface robot for ambient intelligent environments, In Proceedings of the 1st International Workshop on Advances in Service Robotics (ASER), Bardolino, March 13-15, pages 132–139, 2003.
2. D. Fox, W. Burgard, and S. Thrun Markov localization for Mobile Robots in Dynamic Environments, Journal of Artificial Intelligence Research, 11:391–427, 1999.
3. K. Konolige Small Vision System: Hardware and Implementation, In Proceedings of the 8th International Symposium on Robotics Research, Japan, 1997.
4. B.J.A. Kröse, J.M. Porta, K. Crucq, A.J.N van Breemen, M. Nuttin, and E. Demeester Lino, the User-Interface Robot, In First European Symposium on Ambience Intelligence (EUSAI), 2003.
5. B.J.A. Kröse, N. Vlassis, R. Bunschoten, and Y. Motomura A Probabilistic Model for Appearance-based Robot Localization, Image and Vision Computing, 19(6):381–391, April 2001.
6. J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer Multi-Camera Multi-Person Tracking for EasyLiving, In IEEE Workshop on Visual Surveillance, pages 3–10, July 2000.
7. A. Ladd, K. Bekris, G. Marceau, A. Rudys, D. Wallach, and L. Kavraki Using Wireless Internet for Localization, In Proceedings of the International Conference on Robotics and Intelligent Systems (IROS), Las Vegas, USA, pages 402–408, October 2002.
8. H. Murase and S.K. Nayar Visual Learning and Recognition of 3-D Objects from Appearance, International Journal of Computer Vision, 14:5–24, 1995.
9. M.K. Pitt and N. Shephard Filtering Via Simulation: Auxiliary Particle Filters J. Amer. Statist. Assoc., 94(446):590–599, June 1999.
10. J.M. Porta, B. Terwijn, and B.J.A. Kröse Efficient Entropy-Based Action Selection for Appearance-Based Robot Localization In Proceedings of the International Conference on Robotics and Automation (ICRA), Taiwan, 2003.

11. J.M. Porta, J.J. Verbeek, and B.J.A. Kröse Enhancing Appearance-Based Robot Localization Using Non-Dense Disparity Maps In Proceedings of the International Conference on Robotics and Intelligent Systems (IROS), Las Vegas, USA, 2003.
12. N. Vlassis, B. Terwijn, and B.J.A. Kröse Auxiliary Particle Filter Robot Localization from High-Dimensional Sensor Observations In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Washington D.C., USA, pages 7–12, May 2002.