Object Localization using Bearing Only Visual Detection

Kristoffer Sjö AUTHOR [a,1], Chandana Paul AUTHOR [a] and Patric Jensfelt AUTHOR [a]
[a] *Royal Institute of Technology (KTH), Centre for Autonomous Systems, SE-100 44 Stockholm, Sweden*

Abstract.
This work demonstrates how an autonomous robotic platform can use intrinsically noisy, coarse-scale visual methods lacking range information to produce good estimates of the location of objects, by using a map space representation for weighting together multiple observations from different vantage points. As the robot moves through the environment it acquires visual images which are processed by means of a fast but noisy visual detection algorithm that gives bearing only information. The results from the detection are then projected from image space into map space, where data from multiple viewpoints can intrinsically combine to yield an increasingly accurate picture of the location of objects. This method has been implemented and shown to work for object localization on a real robot. It has also been tested extensively in simulation, with systematically varied false positive and false negative detection rates. The results demonstrate that this is a viable method for object localization, even under a wide range of sensor uncertainties.

Keywords. Accumulator Grid, Object Detection, Object Localization

Introduction

Autonomous robots are becoming increasingly free-roaming and independent as the mapping, planning and control systems built into them grow more and more sophisticated. This fact, however, poses a problem in that it increases the rate at which the robot receives new visual data. Visual processing therefore needs to diversify to produce algorithms that are highly reliable but relatively slow on the one hand, and on the other algorithms that are fast but may admit more noise.

In domestic applications many important tasks set for robots involve interacting with specific objects. It thus becomes necessary to develop visual routines for this that allow a robot to make efficient use of its more expensive visual repertoire, by directing attention onto the most likely object locations, and doing this by means of cheaper algorithms.

This paper presents a method for combining relatively low-quality, bearing-only visual object detection output from different vantage points to produce a more accurate estimate of 2D object position, which can then be used for view planning and other probabilistic decision making. It makes use of Receptive Field Cooccurrence Histograms for visual detection and a grid for accumulation of evidence.

[1]Corresponding Author: Kristoffer Sjö, Royal Institute of Technology (KTH), Centre for Autonomous Systems, SE-100 44 Stockholm, Sweden; E-mail: krsj@nada.kth.se

The overall objective is to find objects in the environment and populate the spatial model with them. The work here is a continuation of the work on object search and localization presented in [3,11].

This work is related in purpose, and somewhat in method, to visual attention. There has been a lot of work done in this area, from general saliency-based methods such as that of [9] to more composite approaches such as the one in [7] which is geared specifically towards object detection. These methods operate in the visual space, whereas the approach proposed in this paper instead creates a 2D map representation for the accumulation of evidence.

This 2D map representation has many similarities with occupancy grids [5] where the world is divided into cells. Each cell is assigned a value to reflect the probability or certainty of that cell being occupied. A large number of methods for updating such grids have been proposed, based on Bayesian theory, fuzzy logic, etc. See [16,18] for an overview and comparison of sonar based occupancy grid methods. The sonar sensor provides relatively accurate distance estimates but the angular information is rather vague. Integrating many measurements is therefore necessary to get a good estimate of the structure of the environment. For obstacle avoidance using sonar the Histogram in Motion Mapping approach [1] is sometimes used as it provides a fast way of updating the grid in a time-critical application. In it a simplified update of the occupancy grid is used where only the cells along the acoustic axis of the sensor are updated, by adding and subtracting fixed integer values, thus ignoring the angular uncertainty of the sensor. This is in contrast to the full Bayesian formulation, where in theory, every cell will be affected by every observation. Accuracy in the model is instead gained by frequent updates.

Methods akin to the occupancy grid have also been used for global localization of a mobile robot [6]. Here the grid represents a discretization of the probability density function for the pose of the robot. Recently, particle filters have become popular in robotics applications [2]. A major challenge when estimating a high dimensional state such as the concurrent position of a large number of targets is that of computational cost. In [17] a particle filter implementation of the probability hypothesis density (PHD) is presented. The idea with the PHD is to limit the estimation to the first moment of the state, i.e. the mean value, and also look at the combined density over all targets instead of individual ones. This results in a significant reduction in the computational effort.

As in the case of integrating sonar data in an occupancy grid, each measurement in itself provides relatively little information about the position of an object. Here, the information is bearing-only and depending on the detection method used might have a large amount of false positives. Several measurements are necessary to get a good estimate of the probability distribution. This method is similar to computing the PHD as described above, doing so for each object class independently.

The main inspiration for the way evidence is accumulated in this paper is the Hough transform [8], in that it in effect enumerates possible explanations for observations and accumulates evidence for these explanations. However, the Hough transform involves moving into a configuration space separate from the space being described, which is not the case here.

In object localization itself much work has also been done, including attention and view planning ([15], [14]), as well as detection, distance estimation and localization [3, 10]. The Receptive Field Cooccurrence Histogram (RHCH) method [4] used in [3] is what provides the input for the algorithm presented in this paper.

Finally the area of reliable object recognition has a peripheral bearing on this work, in that recognition will likely be used in the areas indicated by the algorithm proposed; this is however beyond the scope of this paper.

Contributions

The contribution of this paper is an algorithm which allows many individually relatively unreliable sensor responses to be combined efficiently into a map-space representation of confidence values from which more reliable position hypotheses may be extracted. Unlike occupancy grid methods, this confidence estimate uses no range information.

The feasibility of the approach is verified by an implementation running on a real robotic system. Also, results from simulations are presented illustrating the performance of the algorithm and its degree of robustness to false positives and false negatives in the sensor data.

1. Image processing

The algorithm presented in this work is designed to work with Receptive Field Cooccurrence Histograms (RFCH). RFCH is an image processing method for object detection based on bulk image properties, as opposed to feature-based methods such as SIFT [12] [13] which are typically used for recognition. It is used here because it is fast and produces a scalar degree-of-match as output given any part of an image, large or small. RFCH are object-specific and so this algorithm, based on them, will be as well. However, any object detection algorithm with similar properties to RFCH matching could in principle be substituted.

Prior to engaging in the object search, the robot's vision system is trained on each object of interest by performing feature value clustering and histogram extraction on training images of the objects of interest. The clusters and the histogram for the object are stored for the purpose of object detection on the images subsequently acquired.

2. Accumulator Grid

In order to create and maintain a distribution across space of the confidence the robot has in the presence of objects, an *accumulator grid* has been developed. This is a grid overlaid on the robot's navigation map, where each grid cell is associated with a confidence value similar to that of an occupancy grid. One set of such values is maintained for every distinct object of interest. In the following description, only one object is considered for simplicity.

As the robot moves through the environment and acquires visual data, the accumulator grid cells are updated according to the output of the visual object detection algorithm. As data accumulates from different locations, the confidence values reinforce in a way similar to the functioning of the Hough transform, embodying an increasingly accurate representation of the true locations of objects in the environment, and the robot can utilize these to direct its motion, or store them for later use.

2.1. Initialization

The accumulator grid is set up with a specific extent in the X and Y dimensions as well as a cell size. Appropriate values for these parameters are highly dependent on the environment the robot operates in. In general, its extent should equal the size of the operating area and the cell size should be roughly equal to the size of sought objects. If objects of different sizes are to be represented, the smallest size can be used or, alternatively, several grid sizes may be used in parallel.

Initially, the confidence values are all set to 0, if the robot does not possess any prior information about object locations. It would be possible to represent prior knowledge of the likely and unlikely locations of objects during initialization as well. Also, a robot may well be initialized with the values accumulated during an earlier run, if objects are assumed to be static in the interim.

2.2. Update

As the robot proceeds through its surroundings, it acquires camera images, which are subdivided into small patches for which RFCH are extracted. The resulting histograms are compared to the histogram from the training image of the object being sought, producing a match value for each patch.

It is assumed that the pose of the robot is known at all times. If this is not the case, it will become necessary to take steps similar to those needed for occupancy grids built during mapping under uncertain odometry; such considerations are however beyond the scope of this paper.

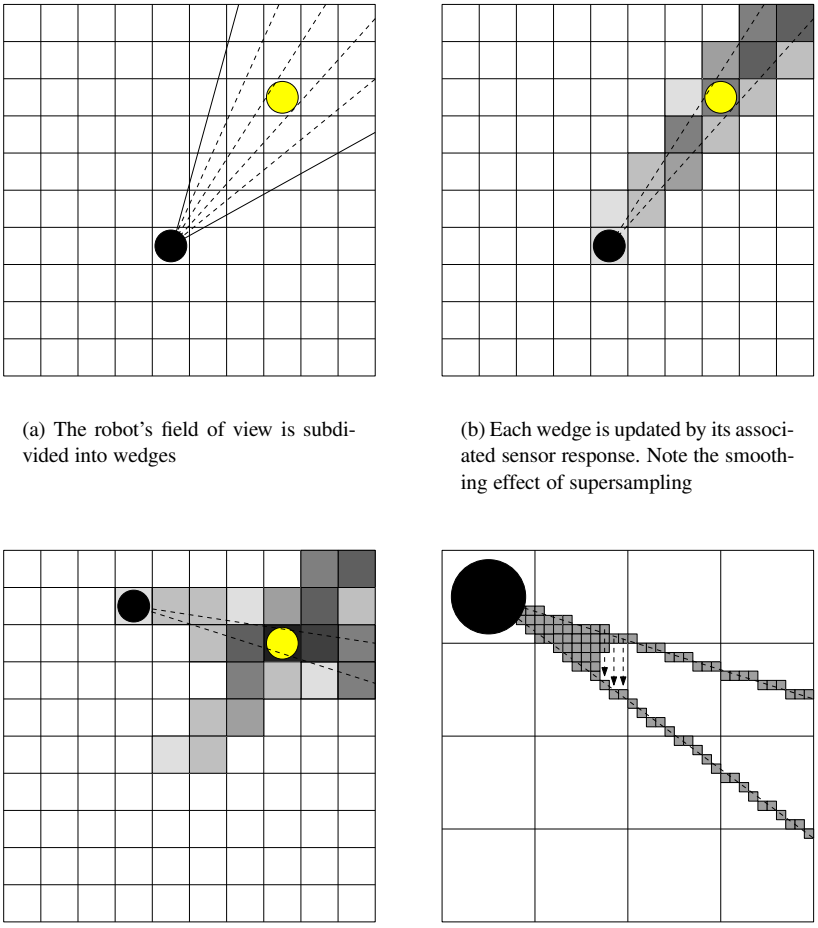
Because the robot's camera is kept horizontal, each column of image patches corresponds to a specific interval of bearings, which is projected onto the 2D map from the location of the robot. This produces a "sensor wedge" as shown in Figure 1(a).

The accumulator grid is then updated according to the following principle: Each cell covered by a sensor wedge is incremented by the maximum magnitude of all the RFCH match values associated with it. In effect, a strong RFCH response will increase confidence in all cells lying in the direction from which that response was obtained, as in Figure 1(b). As the robot continues to acquire images from different vantage points, the wedges intersecting objects will reinforce the grid cells, whereas cells that are only incidentally part of a sensor wedge will not get reinforced as in Figure 1(c).

In practical terms, the update of the cells in each sensor wedge is carried out by means of line rasterization, performed in parallel for the left- and right-hand rays of the slice, respectively, and by filling in the intermediate cells in rows or columns as appropriate. However, given the potentially very large granularity of the accumulator grid and the aliasing effects this will cause, a supersampling scheme is adopted in which the resolution of the grid is augmented by a factor of 10 during the update. In effect, each cell is updated in proportion to how well it is covered; see Figure 1(d). This alleviates aliasing problems without incurring any extra storage costs.

3. Implementation and Experiments

The algorithm described in the previous section was implemented and tested on a Performance Peoplebot mobile robot platform. The robot is approximately 1.2m in height, and



(a) The robot's field of view is subdivided into wedges

(b) Each wedge is updated by its associated sensor response. Note the smoothing effect of supersampling

(c) As views are acquired from different vantage points, cells containing objects will be reinforced

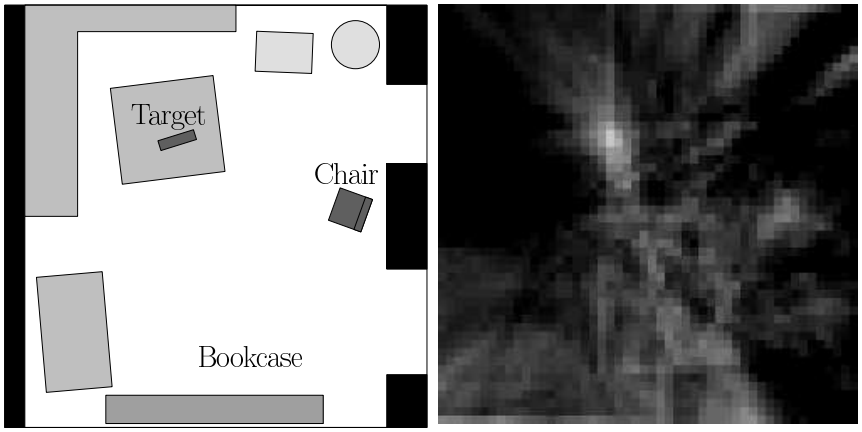
(d) Cells in each wedge are incremented, using supersampling to counter aliasing effects

Figure 1. Principles of the accumulator grid update

equipped with a Canon-VCC4 pan-tilt zoom camera, and differential drive. The camera is mounted at a height of 1m above the floor, and has a horizontal field of view of 45 degrees. The camera was used to acquire images at a resolution of 320×240 pixels. In order for the robot to be able to detect objects in a wide field of view as it explores the environment, the camera's zoom was not used.

An accumulator grid of 50 × 50 cells was created at a resolution of 0.1m, and with all values initially set to zero. The test was performed in a mockup living room of approximate dimensions 4.5m × 6m, shown schematically in Figure 2(a). The robot was programmed to visit 5 locations in the room and take pictures in all directions (8 views at a field-of-view of 45°) from each location. RFCH detection was carried out on the resulting 40 images, and an accumulator grid was updated according to Section 2.2.

After visiting each location and processing the images, the result was the accumulator grid shown in Figure 2(b). The actual location of the sought object, a packet of rice, is plainly visible in the upper left quarter of the grid. On the center right a chair of somewhat similar color to the object can be made out, and towards the bottom of the grid a bookcase containing many items of varying appearance causes a low-level blur; still, these false positives are clearly less prominent than the true location. A view planning algorithm can use this information to create priorities for regions to investigate, which would lead it to begin with the area that actually contains the object in this case.



(a) Schematic representation of the room used

(b) Result of grid accumulation. The true object location is at the bright spot on the upper left

Figure 2. Experimental results

4. Simulation

In order to test the performance limits of the algorithm with respect to visual detection and recognition algorithms which are noisy such as RFCH, SIFT and others, the performance of the system is evaluated by means of Monte Carlo simulation in an abstract scenario, in which the false positive and false negative detection rates of the robot detection can be varied freely.

The environment is represented as a 20×20 grid. The object is positioned randomly in the environment (excepting the outer edges), and occupies exactly one grid cell. The robot can move to any part of the environment except for the cell containing the object. The robot can also have any orientation in space. In the real world, the robot would follow some continuous trajectory through space, sensing as it went. It would thus see the object from several distinct orientations and views. Here, in order to simulate this fact, while avoiding any bias introduced by a manually selected trajectory, the robot is given a random new orientation and position for every view that is acquired.

For a given robot pose, a field of view is simulated by an angular slice of the map, with its apex at the robot's position, as was seen in Figure 1(a). The field of view is divided into wedges corresponding to the image patch columns described in Section 2.2. For each wedge, the cells covered by it are incremented if the object intersects the wedge.

After a set number of random views, the search is terminated and the result is evaluated by locating the maximally-valued cell and comparing its position in the map to the known position of an object. The test is considered successful if the cell with the maximum value in the accumulator grid is the one containing the object or adjacent to it.

The performance of a noisy visual detection algorithm such as RFCH matching was simulated as a detector which gives a binary response on the location of the object, with non-zero false positive and false negative rates P_{fp} and P_{fn} respectively. The performance of object localization with the accumulator grid, under such noisy visual detection conditions, were evaluated in a series of tests. The false positive and false negative rates were varied in intervals of 0.1 between 0 and 1. For each combination of values, tests were performed evaluating the localization with 10, 25, 75 and 150 views. 100 such tests were performed for each parameter setting. Figure 3 shows the outcomes of the tests.

Obviously, a low view count does not provide enough data for a good estimate, but even with 10 random views a good guess can often be made. With 25, results are reasonably reliable in the absence of noise. The real-world test in Section 3 with 40 well-planned views compares to a situation with 75 or 150 random views. At this level the algorithm is able to deal with a large range of false negative rates if the false positive rate is low, and a somewhat more limited range of false positive rates if the false negative rate is low. This is because a false negative does not alter the accumulator grid, whereas a false positive introduces flawed data.

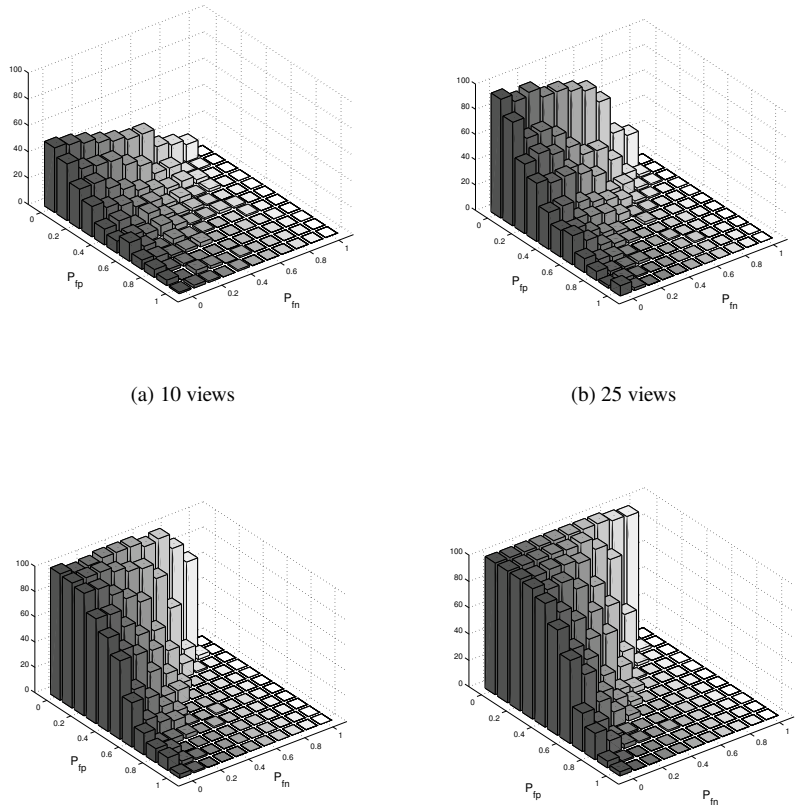
Typically visual processing algorithms are associated with a *Receiver Operating Characteristic* or ROC curve, describing how the false positive rate relates to the false negative rate for that algorithm when varying some discrimination threshold. For RFCH, for instance, a threshold on the degree of match will determine these error rates. Figure 3 shows that good results are achieved whenever either P_{fp} or P_{fn} can be made small, which is the case with many algorithms. More generally, with knowledge of the ROC curve, it is possible to optimize the performance of the accumulator grid for any given algorithm.

5. Discussion

The simulated and real results presented in this paper are promising, and suggest that this algorithm would be a viable method for object localization. However, there are several issues which could be considered further.

Extra-visual information

Information arising from other sources than vision may be useful to take into account. For example, if the layout of walls and other partitions in the environment is known this can be exploited to shield obscured areas from being accumulated needlessly; similarly, objects of interest may be precluded from occurring in certain areas such as in open floor spaces. Prior information on probable locations could also be used.



(a) 10 views

(b) 25 views

(c) 75 views

(d) 150 views

Figure 3. Performance of object localization for various probabilities of false positive and false negative rates. The Z axis denotes the percentage correct estimations, and the X and Y axes the rate of false positives and negatives, respectively.

Viewpoint bias correction

In simulation, the robot was given a new position and orientation at every time step. However, in the real world the robot doesn't move randomly through space, and subsequent locations and views of the object are not entirely independent of previous locations. If there is a bias in the locations from which images are acquired, this will impair the localization.

Extrinsically, this problem can be solved by planning for the robot to move and obtain visual data in a way that provides a good distribution of viewpoints. If this is not possible, intrinsic solutions involving changes to the algorithm itself might be made. Possibilities include weighting of the grid increment based on the amount of sensor data gathered from the current direction, or the creation of multiple grid layers for different viewing directions. Either way would require increasing the amount of data stored.

Usage methods

The intended use for the accumulator grid is to allow a robot to obtain a notion of the likely locations of objects, using low-cost visual procedures, typically as a by-product of other activities such as exploration or more directed tasks that do not monopolize vision. When a new task indicates a need for localization of a given object, the robot will process its accumulator grid for that object, find the likeliest location, and proceed to investigate it, typically using more advanced and expensive visual procedures.

In addition to its primary function of guiding navigation and visual attention, the accumulator grid could also be used for building statistical models of object class distributions in the environment, verbally conveying uncertain data to humans, or possibly performing place classification. The algorithm should be easily usable for all these purposes without requiring any major alterations to its current form.

6. Conclusion

This paper presents a novel method for consolidating noisy bearing-only visual data to achieve object localization. The method uses an accumulator grid to update confidence information through an algorithm that transforms the data from visual into map space. Results are presented of feasibility testing in a realistic scenario as well as an extensive evaluation in simulation. The results indicate that the method performs well in the face of noisy measurements and promises to be useful as a way of obtaining and representing the uncertain location of objects in a robot's environment.

Acknowledgements

This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "CoSy" FP6-004250-IP. Kristoffer Sjö was supported in part by the Swedish Research Council, contract 621-2006-4520.

References

[1] J. Borenstein and Y. Koren. Histogram in-motion mapping for mobile robot obstacle avoidance. *IEEE Transactions on Robotics and Automation*, 7(4):535–539, August 1991.

[2] Frank Dellaert, Wolfram Burgard, Dieter Fox, and Sebastian Thrun. Using the CONDENSATION algorihm for robust, vision-based mobile robot localization. In *Proc. of the IEEE Computer Society Conference of Computer Vision and Pattern Recognition*, volume 2, June 23-25, 1999.

[3] S. Ekvall, D. Kragic, and P. Jensfelt. Object detection and mapping for service robot tasks. *Robotica: International Journal of Information, Education and Research in Robotics and Artificial Intelligence*, 2007.

[4] Staffan Ekvall and Danica Kragic. Receptive field cooccurrence histograms for object detection. In *Proc. of the IEEE/RSJ International Conference on Robotics and Automation (IROS'05)*, 2005.

[5] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.

[6] Dieter Fox, Wolfram Burgard, and Sebastian Thrun. Markov localization for mobile robots in dynamic environments. *Journal of Artificial Intelligence Research*, 11:391–427, 1999.

[7] S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*. PhD thesis, University of Bonn, July 2005.

[8] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, 1988.

[9] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell*, 20(11):1254–1259, 1998.

[10] D. Gálvez López. Combining object recognition and metric mapping for spatial modeling with mobile robots. Master's thesis, Royal Institute of Technology, jul 2007.

[11] D. Gálvez López, K. Sjö, C. Paul, and P. Jensfelt. Hybrid laser and vision based object search and localization. Work presented to the 2008 IEEE International Conference on Robotics and Automation, oct 2007.

[12] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision (ICCV 1999)*, pages 1150–57, Corfu, Greece, September 1999.

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

[14] A. Oliva, A. B. Torralba, M. S. Castelhano, and J. M. Henderson. Top-down control of visual attention in object detection. In *ICIP (1)*, pages 253–256, 2003.

[15] G. Olofsson, J. K. Tsotsos, H. I. Christensen, and S. J. Dickinson. Active object recognition integrating attention and viewpoint control. In *ISRN KTH*, 1994.

[16] M. Ribo and A. Pinz. A comparison of three uncertainty calculi for building sonar-based occupancy grids. In *SIRS*, Coimbra, Portugal, July 1999. A revised version will appear in Journal of Robotics.

[17] Hedvid Sidenbladh. Multi-target particle filtering for the probability hypothesis density. In *In Proc. of the 6th International Conf. on Information Fusion*, pages 800–806, Cairns, Australia, 2003.

[18] Olle Wijk. *Triangulation Based Fusion of Sonar Data with Application in Mobile Robot Mapping and Localization*. PhD thesis, Royal Institue of Technology, Stockholm, Sweden, April 2001.