Experiments in Visual Localisation around Underwater Structures

Stephen Nuske, Jonathan Roberts, David Prasser and Gordon Wyeth

Abstract Localisation of an AUV is challenging and a range of inspection applications require relatively accurate positioning information with respect to submerged structures. We have developed a vision based localisation method that uses a 3D model of the structure to be inspected. The system comprises a monocular vision system, a spotlight and a low-cost IMU. Previous methods that attempt to solve the problem in a similar way try and factor out the effects of lighting. Effects, such as shading on curved surfaces or specular reflections, are heavily dependent on the light direction and are difficult to deal with when using existing techniques. The novelty of our method is that we explicitly model the light source. Results are shown of an implementation on a small AUV in clear water at night.

1 Introduction

We are interested in the localisation of underwater robots around fixed infrastructure. There are many applications of underwater robotics where it is critical for the robot to know where it is with respect to a structure, such as inspection tasks and welding. Assuming that most structures are passive, ie. they do not transmit any location information, then there are two viable sensing modalities that can be used to image a structure; sonar and computer vision. Of these, we have been investigating

Stephen Nuske
The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA e-mail: nuske@cmu.edu

Jonathan Roberts and David Prasser
Autonomous Systems Lab, CSIRO ICT Centre, PO Box 883, Kenmore, Queensland 4069, AUSTRALIA e-mail: jonathan.roberts@csiro.au and david.prasser@csiro.au

Gordon Wyeth
School of Information Technology and Electrical Engineering, University of Queensland, St Lucia, Queensland 4072, AUSTRALIA e-mail: wyeth@itee.uq.edu.au

the use of vision in order to localise an Autonomous Underwater Vehicle (AUV) with respect to a known piece of underwater infrastructure - the leg of a surface platform.

Typically, the visual environment around such a structure is poor. Firstly, suspended particles in the water reduce visibility. Secondly, there is minimal or no natural lighting deep underwater, thus requiring an artificial light source to be mounted on the AUV. Thirdly, the visual appearance of the structure in this scenario is highly dependent on the incident angle of the light source. The light source is constantly moving (as it is on the AUV) and consequently the visual appearance of the structure varies dramatically over time. This is quite different from typical well lit environments, where the light source (typically the Sun) is far less dynamic and also where there is a significant level of ambient lighting. However, rather than this poor visual environment being a negative, we would argue that we can turn it to our advantage. By modeling the light source mounted on the AUV we can predict the appearance of the structure (the legs of a platform in our example) from different viewing poses. The process is to use an a priori 3D-surface model of the permanent structure being navigated with the light model to generate artificial images which are compared against the real camera image to localise the AUV.

2 Previous Work

Kondo et al. present two methods of navigating underwater structures in [10] and [9]. In [10], two laser beams are directed at the structure which are detected in the camera images to triangulate the relative distance and orientation of the vehicle. In [9], Kondo et al. use a light stripe to illuminate a 2D profile of the structure which is detected in the camera images. A common feature of the two systems developed by Kondo et al. is the use of active lighting. In our work an artificial light source is also used, but unlike the focused beams or light stripes of Kondo et al., the light source is unfocused.

Stolkin et al. [13] present work for a submarine localising from a leg of a platform and use an explicit 3D model of the structure, projecting the model onto the image plane to predict the shape of the structure. Model based tracking is an attractive approach for this application as the form of the structure is well known *a priori*. Fig. 1 shows the basic idea behind model based visual tracking. Synthetic images are generated for a large number of possible robot poses and each of these images is compared with the actual image captured by the robot. The comparator can take many forms. Examples include taking the pose that gives the best match, or using a multi-hypothesis framework such as a particle filter [8].

In the wider robotics field, model-based tracking has received much attention. The works of Gerard and Gagalowicz [4], Noyer et al. [12] and Ho and Jarvis [5] present pose estimation systems based on 3D-surface maps. They perform correspondences between real and synthetic images. Both Noyer et al. [12] and Ho and Jarvis [5] estimate pose with a probabilistic particle filter, which is an efficient

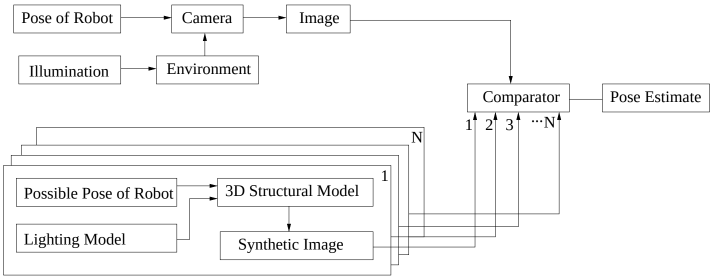

Fig. 1 An overview of the idea of model-based visual pose estimation. Multiple synthetic images (generated at different possible poses of the AUV) are compared with the real camera image of the scene.

means of sampling the solution-space, whereas Gerard and Gagalowicz [4] present a more brute force evaluation of the solution-space. None of these 3D-surface based methods consider reflectance and lighting properties in their work – they only use textured 3D models – which do not generalize to any lighting condition. A textured model would not suffice for the application presented in this paper, because the structure is made of one material and is therefore essentially texture-less. The images of the structure are also highly dependent on the light source, indicating that both reflectance properties of the structure and a light model should be known.

The work of Kee et al. [7] and Blicher et al. [1], in the domain of face identification, introduce the idea of using a 3D-surface model together with a light model. They show how to perform face identification in unknown lighting conditions by first estimating the current light source, then generating synthetic images of each face model using the estimated light source model. They used a database of many different 3D-surface face models. A single fixed pose of the faces with respect to the camera was assumed, then multiple synthetic face hypothesis images were matched to the real image. However, in our work there is a single 3D-surface map of the environment (the structure) and multiple pose hypotheses that are matched to a real image (taken from the AUV). The pose hypotheses with the best image match to the camera image from the robot will provide the pose estimate. This idea of estimating and incorporating a light model has not yet been applied to visual localisation.

3 Localisation Framework

Our framework uses a model of the structure and a model of light source together to generate synthetic images that are expectations of the real camera images. The synthetic images are compared to the real images to estimate the pose of the camera. The pose estimation is facilitated in a probabilistic multiple pose hypothesis framework – a particle filter – which uses a synthetic image of the structure from each pose hypothesis to derive a comparison score against the real image.

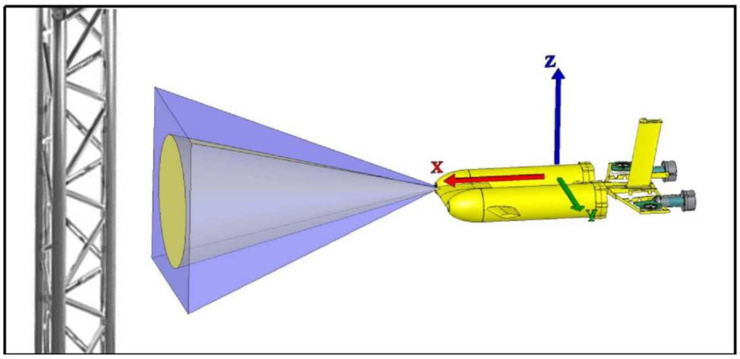

Fig. 2 The AUV has a forward facing camera with a field-of-view depicted in the figure by pyramid viewing volume. The spotlight is also facing forward and partially illuminates the field-of-view of the camera (depicted by the inner cone).

A forward looking spotlight and camera are mounted rigidly to the AUV[3] which is inspecting a structure. Fig. 2 shows the AUV, the coordinate system, the camera and spotlight-setup. A single forward-facing camera, from the AUV's stereo pair, is used. The field-of-view of the camera, is shown as pyramidal viewing volume in Fig. 2. The extrinsic pose of the camera is calculated with respect to the vehicle, and is facing along the vehicle's positive x axis. The camera's intrinsic parameters were calibrated using the OpenCV library[6]. The camera images are then undistorted by these parameters, and the model of the structure can then be projected directly onto the image plane.

3.1 Synthetic Image Generation

The synthetic images are generated from a polygon mesh of the structure. The mesh includes the surface normal, diffuse and specular reflectance properties. Meshes using such detailed surface properties have rarely been applied to visual localisation. These photometric properties are incorporated into the Blinn-Phong model [2] using the OpenGL library for generating synthetic images. There are other lighting models which could be also used, but the Blinn-Phong model is chosen for its simplicity, speed of computation and prevalent implementation on most graphics processors. In addition to the pose of the light, the model incorporates a number of other parameters to account for the attenuation by water and angular spread of the light source.

The structure to be localised against is comprised of three steel tubes, linked together by smaller rungs at approximately 45 degree angles. This structure is modelled as a polygon mesh of each of the tubes by defining the number of slices and stacks in the mesh of each tube. The rendering is performed with per-fragment lighting calculations, and shading interpolation to the pixels within the fragment, according to the light model defined above. A series of tests were performed to calculate the optimum number of polygons taking into account render times and model accuracy. For this model and using an Intel dual core 2.33GHz CPU and a NVIDIA Quadro FX 350M GPU with 320x240 images, it was found that optimum value of 3120 polygons equated to a time of 0.325ms to render a single synthetic image.

3.2 Particle Filter Localisation

The use of a particle filter is described in detail by Thrun et al. in [14]. The particle filter is a set of N pose hypotheses (particles) $X_t = x_t^{(1)}, x_t^{(2)}, x_t^{(3)} \ldots, x_t^{(N)}$. The set is sampled from the previous set X_{t-1} using a propagation model m_t and a corresponding set of weights (probabilities), W. The weights are calculated from an observation of the environment, y, as follows:

$$W_k^{(n)} = p(y_k | x_k^{(n)}) \tag{1}$$

the observation of the environment is a camera image, y, which is compared with each pose particle x by rendering a synthetic image. The measurement of probability is provided from an image matching technique (discussed in Section 3.3). The concept is that a synthetic image generated from the particles nearest the correct pose will give the best match with the real image. These particles are then the most likely to be re-sampled for the next iteration. The current pose estimate of the AUV is extracted from the filter as the mean pose of the particles with the highest weights (top 5%). We use roll and pitch estimates from the AUV's Inertial Measurement Unit (IMU) as each particle's roll and pitch estimate. The remaining four degrees of freedom are propagated using a constant velocity model calculated from a set of previous pose estimates extracted from the particle filter.

3.3 Gradient-domain Image Matching

The comparison process between the camera image and a synthetic image provides a likelihood measure for the set of particles in the filter. The works of [1, 4, 7, 12], all use image matching techniques which compare real and synthetic images. All of the techniques are variants of the Mean Absolute Difference (MAD).

The simple image matching techniques assume that it is possible to generate pixel intensities for the synthetic image that are equivalent to those in the real image. This is different from photo-realistic rendering, which is only interested in making the synthetic image appear real. Whereas, these image intensity matching techniques require the environment model and light model to be accurate estimates of the actual physical properties of the surrounding environment. The parameters of such models are difficult to estimate accurately. Furthermore, the Blinn-Phong image formation model [2], employed for real-time performance, does not incorporate the ability to model poor visibility. Suspended tiny particles that cause poor visibility have two effects on the lighting; absorbing light and reflecting light. These effects would need to be modelled before accurate intensity values could be generated in the synthetic image. It is difficult to generate accurate images of simulated poor visibility conditions at a rate quick enough for this framework. For this reason, intensity-based matching is not used and, instead, a gradient-domain image matching technique is developed. The gradient-domain removes the absolute intensity levels whilst capturing the subtle shading in the environment. This behaviour is different to an edge-image that identifies drastic boundaries of intensity.

The first step is to pass the real-image through a Gaussian filter, which removes the effects of noise. The synthetic and real images are then both passed through a horizontal Sobel operator to generate gradient images in both the x and y directions; G_x and G_y are the real Sobel images and g_x and g_y are the synthetic. Example synthetic and real images are shown in Fig. 3. The x direction is shown in red and y in green, therefore pixels with high x and y gradients are yellow. To compare the real and synthetic Sobel images, it would be possible to turn these two images into a gradient magnitude image and a gradient orientation image, which would enable a more logical means for comparison. But to avoid the expensive square root and arc tan computations, the images are compared directly in x and y gradients.

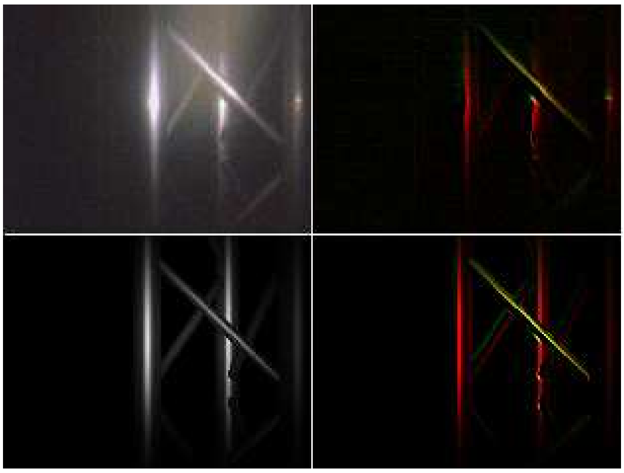

Fig. 3 Top Left: Real camera image. Top Right: Real Sobel image. Bottom Left: Synthetic image. Bottom Right: Synthetic Sobel Image. Horizontal gradient is shown in red and vertical in green; therefore pixels with diagonal gradients are yellow.

Firstly, a sum is taken of the gradient magnitude in the real, S_r, and synthetic, S_s, images:

$$S_r = \sum_{p=0}^{N} (|G_x(p)| + |G_y(p)|) \qquad S_s = \sum_{p=0}^{N} (|g_x(p)| + |g_y(p)|) \tag{2}$$

where N is the number of pixels. Secondly, a sum of the difference in gradients between real and synthetic is calculated in each direction, D_x, D_y;

$$D_x = \sum_{p=0}^{N} (|G_x(p) - g_x(p)|) \qquad D_y = \sum_{p=0}^{N} (|G_y(p) - g_y(p)|) \tag{3}$$

the final image matching score is derived by subtracting the sum of the gradient difference from the sum of the gradient magnitude and normalizing by the sum of the gradient magnitude:

$$D_\mu = \frac{(S_r + S_s) - (D_x + D_y)}{S_r + S_s} \tag{4}$$

This result equates to the observation y and the particle $x^{(n)}$ from (1). The better the match between the images the larger value of D_μ. This score can be incorporated into the observation y and the particle $x^{(n)}$ from (1) as follows:

$$W_k^{(n)} = p(y_k | x_k^{(n)}) \propto e^{\rho D_\mu} \tag{5}$$

Where ρ is a positive constant that adjusts the convergence of the particle filter.

4 Results

An experiment was conducted at night in clear water with the aim of determining if the visual localisation system in combination with the IMU could localise the AUV as it moved freely in all six degrees of freedom. The experiment began with the AUV approximately 1.5m away from the structure. The AUV approached the structure, strafed side to side, descended and rotated around the structure. Note that there was no ground truth data available during this experiment and hence the performance of the system could only be checked manually by inspecting the projected centre lines of the structure from the estimated pose, and confirming they align correctly in the raw camera images. Tracking images from the experiment are presented in Fig. 4, along with a movie of the results can be found in the video attachment located at:

http://www.cat.csiro.au/ict/download/nuske/auv_pooltest1.mpg

The visual localisation system maintained accurate track of the structure for 440 frames where there were significant changes in scale, orientation and translation. The system then made a mistake when one of the columns of the structure disappeared behind another, and then reappeared on the other side. The system estimated the column reappearing on the same side, and did not recover from this error. The frames just before and just after the disappearing column can be seen as the bottom two images of Fig. 4. When the system was run again, and again over the same data, it did occasionally correctly estimate that the rear column appeared on the other side. However, it failed more times than it succeeded and a solution to this problem is currently being investigated.

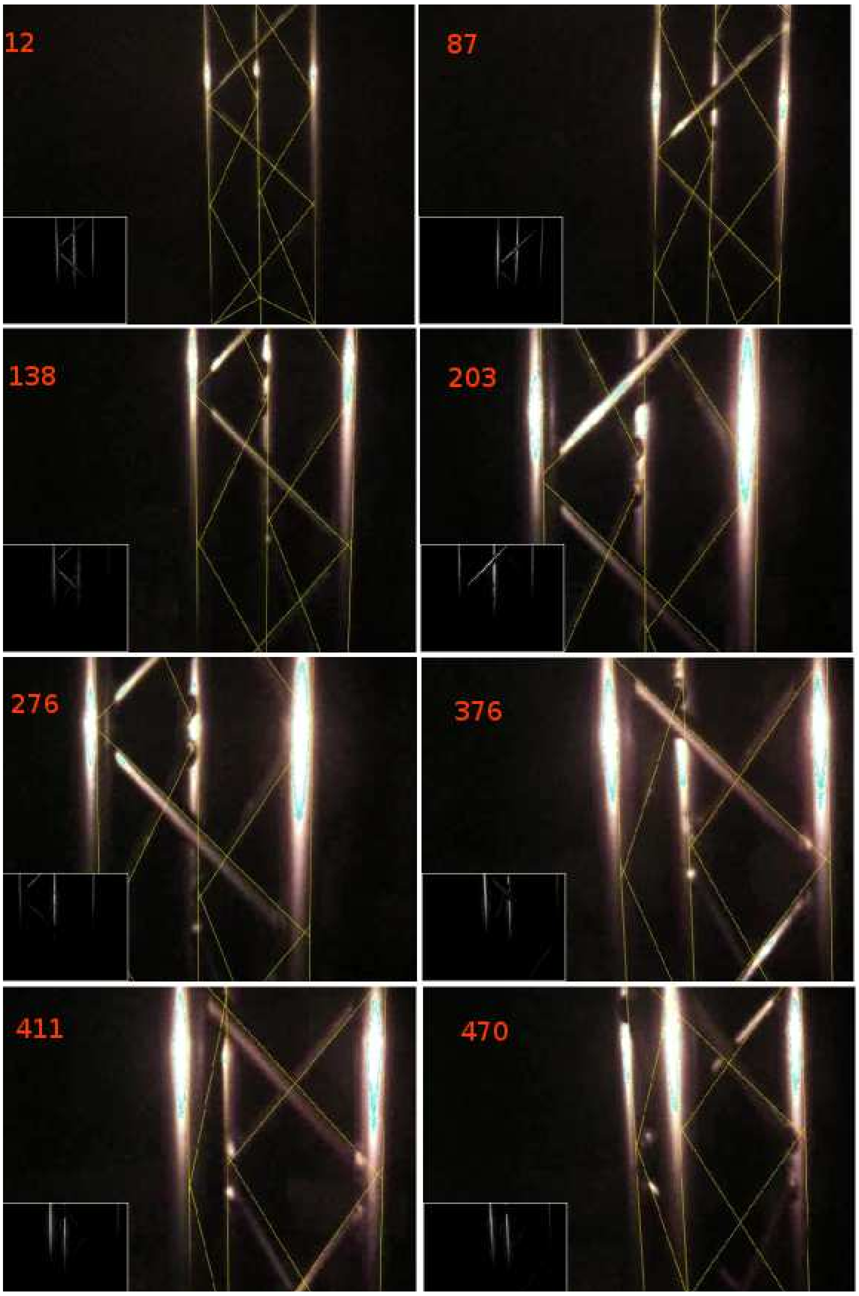

Fig. 4 Images showing the tracking of the oil rig structure. Real camera image is overlaid with the centre lines of the structure projected from the estimated pose. Bottom left corner of each image is a synthetic rendering of the structure. The frame number is located in the top left corner of each image. Data was collected at 15Hz.

Algorithmic speed and efficiency are critical to ensure a practical system for the target application. There is a demand to use low power processors on AUVs due to the on board power limitations of the vehicles. The results reported above were processed off-line. However, the algorithm has been implemented in a such a way as to make it amenable to on-board impementation. The resolution of the 3D-surface was minimised, and the image processing and matching algorithms were implemented on a GPU, which all brought the processing times much closer to real-time rates. The system was run on a laptop computer, which used small mobile processors and graphics hardware. It achieved a frame rate of 0.5Hz with 800 particles using 320x240 sized images and would therefore require the AUV to be moving slowly with respect to the structure to be able to track it reliably in real time.

The two potential methods to achieve higher efficiencies, would be to further reduce the render times of the synthetic images, and also to reduce the number of particles in the filter. Reducing the render time could involve further reductions in the polygon counts, only passing sections of the model that are in view to the graphics pipeline, optimising the lighting calculations or with improved/multiple GPUs. Future improvements to reduce the number of particles could include using a two-stage coarse-to-fine particle filter, such as used in the work of Klein and Murray [8], or to develop a better propagation model. The depth sensor and the magnetic compass are two sensors which could be included in the propagation model. However, it would need to be confirmed that these sensors are locally consistent in the desired environment (that is, if their inter-frame motion estimates are accurate). Another possible method of improving the propagation model is to use a visual odometry system. Marchand et al. [11] present such an approach.

5 Conclusion and Future Work

We have presented a visual localisation system that explicitly models the spotlight of an AUV navigating underwater structures. The light model is used in conjunction with a surface model of the structure to generate synthetic images that are accurate representations of the real camera image. A particle filter framework is employed where a synthetic image is rendered from each pose hypothesis and a observation function computes a probability through comparison with the real camera image. The observation function compares the real image with the synthetic images operates in the intensity gradient domain, avoiding the need to generate precise intensity values in the synthetic image and allows the system to operate in poor visibility conditions which are difficult to replicate in the synthetic image. The system was tested using a monocular vision system, spotlight, steel structure and low-cost IMU. Results show that the system can localise the vehicle in challenging image sequences where the light source is constantly moving and illuminating the scene non-uniformly.

In future work the system will continue to be developed with the goal of a fully functioning system in the targeted offshore environments. Localising from structures of other shapes and surface characteristics will be evaluated. The image processing algorithms presented here are essentially generic and are expected to be able to provide similar results from other structures. Improvements to the lighting model will also be investigated, such as modelling the spotlight as an area light source and also using a more accurate model of the surface reflectance properties to generate images with more precise representation of the shading. More accurate odometry information will also be employed which may come from a compass, a pressure (depth) sensor or potentially a visual odometry algorithm. This information is expected to greatly improve the accuracy and computational efficiency of the system by significantly reducing the area of the state space that must be evaluated. Ambiguous visual scenarios which have been the cause of divergence in the localisation filter will also be investigated in more detail.

References

1. Blicher, A.P., Roy, S., Penev, P.S.: Lightsphere: Fast lighting compensation for matching a 2d image to a 3d model. In: 17th International Conference on Pattern Recognition (ICPR'04), pp. 157–162. IEEE Computer Society, Washington, DC, USA (2004)
2. Blinn, J.F.: Models of light reflection for computer synthesized pictures. SIGGRAPH Comput. Graph. **11**(2), 192–198 (1977)
3. Dunbabin, M., Roberts, J., Usher, K., Winstanley, G., Corke, P.: A hybrid AUV design for shallow water reef navigation. In: Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on, pp. 2105–2110 (2005)
4. Gerard, P., Gagalowicz, A.: Three dimensonal model-based tracking using texture learning and matching. Pattern Recognition Letters **21**, 1095–1103 (2000)
5. Ho, N., Jarvis, R.: Global localisation in real and cyber worlds using vision. In: Australasian Conference on Robotics and Automation (2007)
6. Intel: Open Source Computer Vision Library: Reference Manual (2000). URL http://www.intel.com/technology/computing/opencv
7. Kee, S.C., Lee, K.M., Lee, S.U.: Illumination invariant face recognition using photometric stereo. In: IEICE Trans. on Information and Systems, vol. 7, pp. 1466–1474 (2000)
8. Klein, G., Murray, D.: Full-3d edge tracking with a particle filter. In: British Machine Vision Conference, pp. 1119–1128 (2006)
9. Kondo, H., Maki, T., Ura, T., Nose, Y., Sakamaki, T., Inaishi, M.: Relative navigation of an autonomous underwater vehicle using a light-section profiling system. In: Proceedings of the 2005 IEEE International Conference on Intelligent Robots and Systems (IROS)., pp. 1103–1108 (2004)
10. Kondo, H., Ura, T., Nose, Y., Akizono, J., Sakai, H.: Visual investigation of underwater structures by the AUV and sea trials. In: OCEANS 2003. Proceedings, vol. 1, pp. 340–345 Vol.1 (2003)
11. Marchand, E., Bouthemy, P., Chaumette, F.: A 2d-3d model-based approach to real-time visual tracking. Image and Vision Computing **19**(13), 941–955 (2001)
12. Noyer, J., Lanvin, P., Benjelloun, M.: Model-based tracking of 3d objects based on a sequential monte-carlo method. In: Conference on Signals, Systems and Computers, vol. 2, pp. 1744–1748 (2004)
13. Stolkin, R., Hodgetts, M., Greig, A.: An EM/E-MRF strategy for underwater navigation. In: Proceedings of the British Machine Vision Conference, pp. 715–724 (2000)
14. Thrun, S., Burgard, W., Fox, D.: Probabalistic Robotics. The MIT Press (2005)