

3-D Scene Reconstruction from Cylindrical Panoramic Images

Roland Bunschoten*

Ben Kröse**

RWCP, Autonomous Learning Functions Lab. SNN
Computer Science Institute, Faculty of Science
University of Amsterdam, The Netherlands

Abstract. In this paper we address the problem of recovering 3-D scene structure from omni-directional cylindrical panoramic images obtained from camera viewpoints whose relative poses are known. Range can be estimated using a stereo vision technique from two images. However, range cannot be estimated reliably in the heading direction relating two viewpoints. Moreover, range estimates obtained from a single pair of images tend to be noisy. To overcome these problems we have developed a method which uses multiple images obtained from non co-linear viewpoints in order to obtain more accurate and robust range estimates. Range data obtained from different pairs of images is fused in a probabilistic framework. We present our method and experimental results obtained using image data acquired by our robot. The results indicate that robust and reliable range estimates can be obtained using the proposed method.

1 Introduction

For global, goal-directed navigation a mobile robot needs an internal representation of its environment. Traditionally, the environment is modeled as an explicit geometric model. Recently, several *appearance based* modeling approaches have been proposed, using an (omni-directional) vision system on the robot. In these approaches the relationship between images (or features derived from them) and robot positions is modeled [9].

A drawback of appearance based environment representations is that many training images are needed to obtain an accurate model. An approach to overcome this problem was presented in [3], where based on a number of measured range profiles synthetic profiles are generated. We are working on a similar method to generate synthetic *images* from measured images. In order to do so, dense and accurate range information is required. In this paper we describe an *omni-directional scene reconstruction algorithm* which provides the range estimates required to generate synthetic panoramic images from measured panoramic images.

In earlier work [2] we have shown that range estimates can be estimated from *two* cylindrical panoramic images obtained by an omni-directional camera mounted on top of a mobile robot. Since range is estimated from only two images, erroneous estimates occur due to perceptual aliasing and other phenomena. Also, because of the large field of view covered by the images and the restrictions on camera movement the focus of contraction (FOC) and focus of expansion (FOE) are always visible in the image domain. Range estimates in these directions *cannot* be estimated reliably. The solution to the aforementioned problems we describe in this paper uses multiple images obtained from various non co-linear poses and fuses the range measurements in a probabilistic fashion.

2 Panoramic Stereo Vision

Stereo vision is the process of recovering range information from two images obtained from different but known relative camera poses. The range to a physical point can be computed by triangulation if projections of the point in two images are known (reconstruction problem). Establishing the matching image coordinates is the fundamental problem in stereo vision (correspondence problem).

If knowledge about the camera geometry and relative viewpoints is available, a powerful geometric constraint, known as the *epipolar constraint*, reduces the search space for possible matches from two dimensions

* Email: bunschot@science.uva.nl

** Email: kroese@science.uva.nl

(the entire image plane) to one dimension (the *epipolar curve*). Besides computational gain, an important advantage of exploiting the epipolar constraint is that the likelihood of establishing erroneous pixel correspondences is reduced.

The epipolar geometry relating two viewpoints v^0 and v^1 can be formalized as follows. Let \mathbf{x}^i denote the 3-D position of a scene point expressed in the i -th viewpoints coordinate frame. The transformation between v^0 and v^1 is given by

$$\mathbf{x}^1 = \mathbf{R}^1 \mathbf{x}^0 + \mathbf{t}^1. \quad (1)$$

where \mathbf{R}^1 is a (3×3) rotation matrix and \mathbf{t}^1 is a (3×1) translation vector. Let \mathbf{y}^i denote the projection of \mathbf{x}^i onto the camera image surface. The plane spanned by \mathbf{t}^1 and \mathbf{y}^0 defined by its normal \mathbf{n}^0 is called the *epipolar plane*. In the v^1 coordinate frame the normal to the epipolar plane is computed as

$$\mathbf{n}^1 = \mathbf{R}^1 \mathbf{n}^0 = \mathbf{R}^1 (\mathbf{t}^1 \times \mathbf{y}^0), \quad (2)$$

where \times denotes the outer product. Equation 2 establishes the epipolar geometry; in order to find the \mathbf{y}^1 corresponding to \mathbf{y}^0 only those \mathbf{y}^1 's satisfying

$$\mathbf{n}^1 \cdot \mathbf{y}^1 = 0 \quad (3)$$

need to be searched, *i.e.* the intersection of the epipolar plane with the imaging surface. For planar imaging surfaces this intersection is a line referred to as the epipolar line.

We deal with cylindrical panoramic images. Such images are formed by a perspective projection of the environment onto a unit radius cylinder. A point on the cylindrical imaging surface (*viz.* a pixel) can be characterized by (ϕ, z) (see figure 1). For cylindrical panoramic images the intersection of the epipolar plane with the imaging surface forms a curve in the image domain. The epipolar curve is a sinusoid as can be seen by expressing \mathbf{y}^1 in equation 3 in cylindrical coordinates and by expanding the dot product which gives

$$\mathbf{n}_x^1 \cos(\phi^1) + \mathbf{n}_y^1 \sin(\phi^1) + \mathbf{n}_z^1 z^1 = 0, \quad (4)$$

where subscripts denote components of the normal vector. Rewriting this (dropping superscripts gives

$$z(\phi) = -\frac{\mathbf{n}_x \cos(\phi) + \mathbf{n}_y \sin(\phi)}{\mathbf{n}_z}. \quad (5)$$

The epipolar geometry for cylindrical panoramic images is illustrated in figure 1. In the figure, d_{\min} and d_{∞} represent vectors in the direction of \mathbf{y}_1 whose length is indicated by the subscript.

In order to find the corresponding point in image I^1 for a given point (ϕ^0, z^0) in image I^0 , the epipolar curve can be sampled at equidistant ϕ^1 's resulting in a set of points $(\phi^1, z(\phi^1))$. Matching is done by computing an image similarity measure between windows centered at (ϕ^0, z^0) and the sampled points $(\phi^1, z(\phi^1))$. In a basic setting, the point $(\phi_*^1, z(\phi_*^1))$ which is most similar to the point (ϕ^0, z^0) is selected as the corresponding point. Range can then be estimated from the corresponding points by triangulation (see [2] for details).

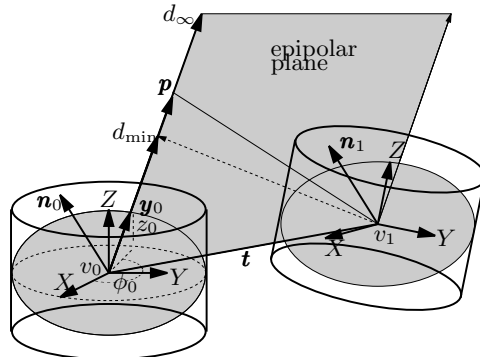


Fig. 1. Epipolar geometry for omni-directional cylindrical images.

3 Scene Reconstruction Algorithm

3.1 Notation and Definitions

We define a reference viewing direction ray as the line through the origin of v^0 and an image location (ϕ, z) in a reference image I^0 obtained at v^0 . Let $p = (\phi, z)_r$ denote the location of a scene point along the reference viewing direction ray. Let $\mathbf{x}^k(p)$ denote the location of p in the k -th viewpoint. Let $\mathbf{y}^k(p)$ denote the projection (expressed in cylindrical coordinates) of $\mathbf{x}^k(p)$ onto the I^k image plane in the k -th viewpoint. Specifically,

$$\mathbf{x}^0(p) = r[\cos \phi, \sin \phi, z]^T$$

and

$$\mathbf{y}^0(p) = [1, \phi, z].$$

The relationship between $\mathbf{x}^k(p)$ and $\mathbf{x}^0(p)$ is given by

$$\mathbf{x}^k(p) = \mathbf{R}^k \mathbf{x}^0(p) + \mathbf{t}^k,$$

where \mathbf{R}^k is a (3×3) rotation matrix and \mathbf{t}^k is a (3×1) translation vector relating viewpoint v^0 and v^k . The projection $\mathbf{y}^k(p)$ is computed from $\mathbf{x}^k(p)$ as

$$\mathbf{y}^k(p) = \left[1, \tan^{-1} \left(\frac{\mathbf{x}_y^k(p)}{\mathbf{x}_x^k(p)} \right), \frac{\mathbf{x}_z^k(p)}{\sqrt{(\mathbf{x}_x^k(p))^2 + (\mathbf{x}_y^k(p))^2}} \right].$$

3.2 Motivation and General Approach

Scene reconstruction from two cylindrical panoramic images is limited by the fact that the FOE and FOC (referred to as epipoles in the context of stereo vision) are always visible in the image domain. In these directions, range *cannot* be estimated reliably. Furthermore, erroneous range estimates can occur due to perceptual aliasing and other phenomena. These problems can be tackled by fusing multiple scene reconstructions computed from multiple image pairs. Note that the first problem is only solved when the various images are obtained from non co-linear poses since images obtained from co-linear poses share the same epipoles.

Sensor data fusion is most convenient when the data to be fused is aligned and expressed in a common frame of reference. In the case of stereo vision, this is related to the way epipolar curves are sampled as follows. Every sampling of an epipolar curve in the image domain amounts to some sampling along a viewing direction ray in scene space. The actual sampling of points along the viewing direction ray differs when corresponding epipolar curves in images obtained from different viewpoints are traversed. The consequence is that in order to be able to fuse range estimates obtained from different image pairs re-sampling of estimated range data is required. Not only does this introduce additional computational overhead but it might also introduce additional errors. The problem is illustrated in figure 2a.

The solution we propose here circumvents the problem of re-sampling altogether. Instead of sampling an epipolar curve according to some image space parameterization as is conventional in stereo algorithms, we propose to *sample the epipolar curve according to a scene space parameterization*. This is done by explicitly generating points, or samples, $\mathbf{x}^0(p)$ along a reference viewing direction ray in 3-D scene space and by projecting each sample to images obtained from other viewpoints v^1, \dots, v^K . Image similarity is computed by comparing small windows centered at the image projections $\mathbf{y}^0(p)$ and $\mathbf{y}^k(p)$. This strategy is illustrated in figure 2b.

3.3 Data Fusion

In our current implementation of this idea a set of N samples $P = \{p_1, \dots, p_N\} = \{(\phi, z)_{r(1)}, \dots, (\phi, z)_{r(N)}\}$, where $r(i)$ is obtained by equidistant sampling from an interval $[r_{\min}, r_{\max}]$, is generated for a selected subset of pixels in a panoramic reference image. Image similarity between I^0 and I^k is computed for each $p \in P$ by evaluating the summed squared difference (SSD) between small square windows taken from I^0 and I^k centered at $\mathbf{y}^0(p)$ and $\mathbf{y}^k(p)$ respectively.

For each sample p_i along a viewing direction of the reference image a set of SSD values is thus acquired. The problem is now to decide whether there is an object at p_i , given the measured SSD values. One approach

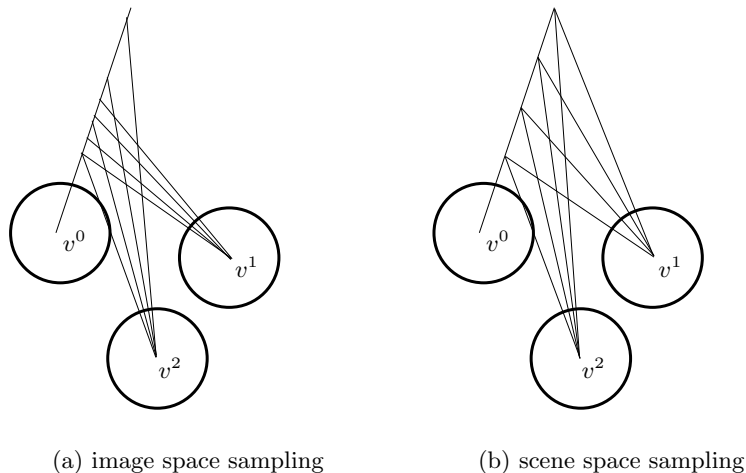


Fig. 2. Sampling the epipolar line from image space and from scene space.

for data fusion is probabilistic fusion, where the probability of an object at p_i is calculated from the measurements. Thus, a model for $P(p_i = occ | SSD^1, SSD^2, \dots, SSD^K)$ is needed. An overview of various probabilistic fusion rules is given in [4]. For the moment we adopt the ‘independent opinion pool model’ and write:

$$\begin{aligned}
 P(p_i = occ | SSD^1, SSD^2, \dots, SSD^K) = \\
 P(p_i = occ)^{-(K-1)} \prod_{k=1}^K P(p_i = occ | SSD^k).
 \end{aligned} \tag{6}$$

where $P(p_i = occ)$ is the prior probability of being occupied.

The individual sensor model is obtained experimentally. For a large number of corresponding windows (windows which observe the same point in scene space) we calculated the SSD. The distribution $P(SSD | p = occ)$ of the data could well be modeled using an exponential function. Therefore we adopt this type of distribution. From this it is easy to see that we have to add the SSD values obtained from every image for a particular p_i for a probabilistic fusion.

For each viewing direction a set of N summed SSD values is available. When asked to return a specific range estimate for a particular viewing direction, the point p for which the summed SSD value is minimal is returned. In the absence of any prior information (*e.g.* assumptions about scene continuity) this corresponds to the maximum likelihood estimate of the physical location of point p given the measurements. Situations where a reference viewing ray is only partially visible from v^k sometimes occur. Since the SSD can only be evaluated when a point is visible both from v^0 and v^k the samples not visible from both viewpoints are ignored.

4 Experiments

Our experimental platform is a Nomad Scout robot (manufactured by Nomadic Technologies, Inc.) equipped with (among other sensors) odometry sensors and an omni-directional vision sensor. The omni-directional vision sensor consists of a vertically oriented camera (Sony EVI-370) and a hyperbolic mirror (manufactured by Accowle, Co., LTD) mounted in front of the camera lens. Using a parameterized geometric image formation model hyperboloid omni-directional images are transformed into cylindrical panoramic images. The mirror model parameters are known a priori from the manufacturers specification. The camera model parameters (such as focal length and lens distortion) are estimated using the camera calibration procedure proposed by Heikkila [6].

Range estimation requires that the relative poses from which images are obtained are known with high accuracy. Although our robot is equipped with fairly reliable odometry, we have noticed that small errors in the robot orientation estimate quickly cause epipolar curves not to pass through corresponding image points.

Therefore, we estimate the transformation between viewpoints from image correspondences. In the experiments reported in this paper these correspondences were manually selected. More recently we have use the Shi-Tomasi feature tracker [11] to provide an initial set of corresponding points. In order to estimate view-point transformation parameters we use an iteratively re-weighted least squares estimation procedure [13]. The estimated parameters are decomposed into a rotation matrix and a translation vector required by our method using an SVD based algorithm outlined in [12]. It is well known that the translation between two viewpoints can only be estimated up to an unknown scale factor from images. We use robot odometry measurements to provide the scale factor.

Figure 3 sketches the layout of the end of a hallway in our building and 5 positions where (360×100 pixels) cylindrical panoramic images were acquired. A reference image was acquired at the center of the depicted cross. The other images were acquired at the ends of the cross line segments. The acquired images are displayed in figure 3. In the experiments the upper image is regarded as the reference image.

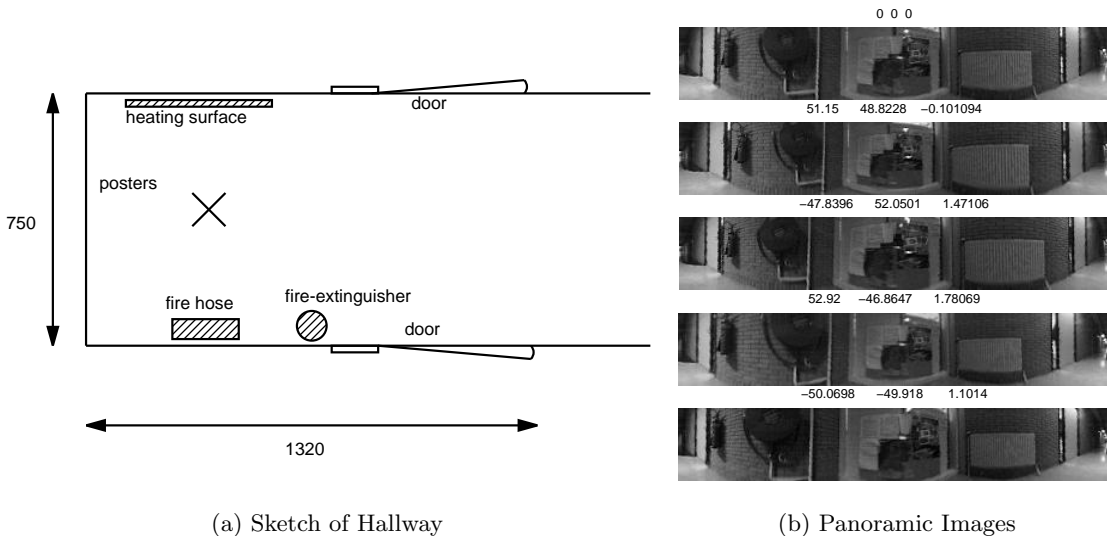


Fig. 3. Outline of the environment and panoramic images.

Figures 4a–d display local range maps obtained by our method using the reference image and the four images obtained at each of the other viewpoints. For visualization, for every viewing direction only the point $p \in P$ for which the sum of SSD values is minimal is displayed. Note that a maximum likelihood range estimate is provided for every viewing direction; no reliability measure has been used to reject unreliable estimates.

Figure 5a presents the range map obtained after fusing all local maps. The resulting range map is clearly better than any of the individual maps. We have investigated the outliers that remain and found that these are caused by perceptual aliasing (heating apparatus and doors) and by specular reflections (posters). Figure 5b presents a top view of the range estimates (for viewing directions parallel to the floor only). From this plot it can be seen that the scale at which range is recovered is correct.

5 Discussion

In this section we will discuss several extensions and enhancements of the current implementation. Memory usage can be reduced in two ways without affecting the accuracy or representational power of the environment model build much. First, instead of computing a local environment map for each of the K images and fusing results afterwards, fusion can be done in an iterative manner. That is, a local map is computed from image I^0 and I^1 and the local map resulting from image I^0 and I^k can be directly integrated with the map already available without explicitly storing the new map. Secondly, instead of retaining all samples along a viewing direction, only those samples more likely to correspond to a physical scene point (based on an already

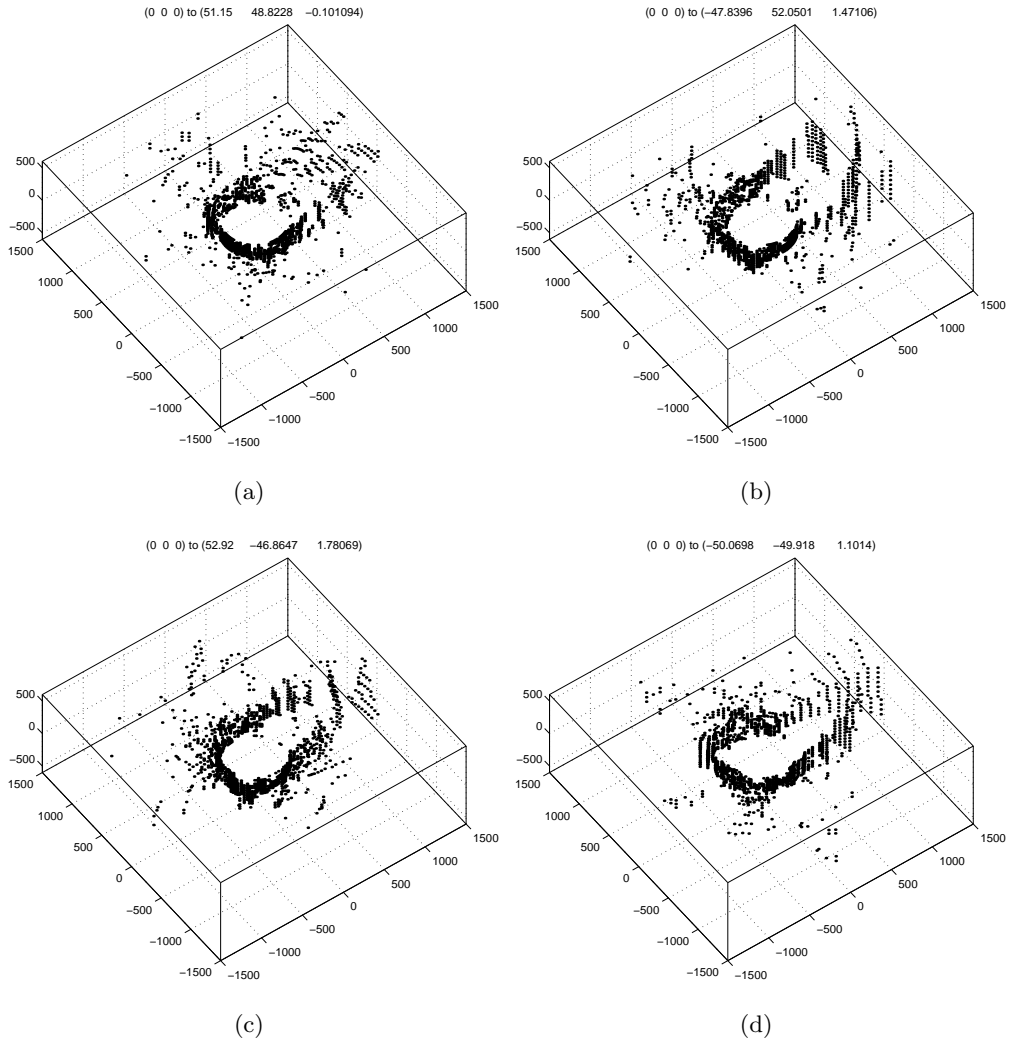


Fig. 4. Range estimated from the reference image and the images obtained at other viewpoints.

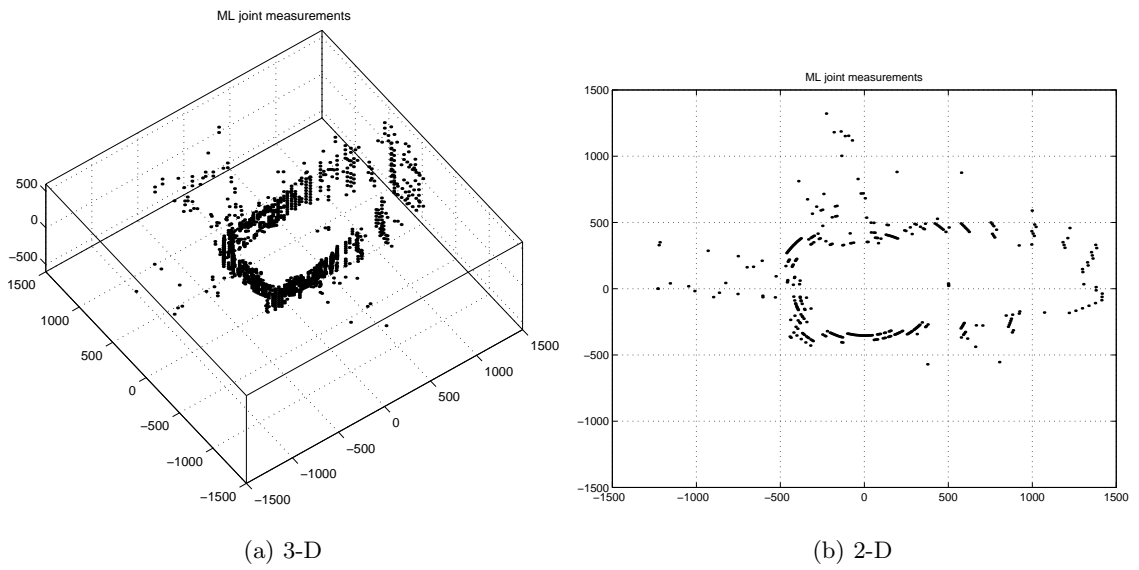


Fig. 5. Range estimated by fusing information obtained from all viewpoints

established map) can be kept. This idea is similar to conditional density propagation (aka condensation) which has been successfully applied in the field of visual object tracking [1] and mobile robot localization [5].

Further planned extensions include the incorporation of explicit assumptions on the shape of the environment (*e.g.* that the environment is locally smooth, or that an indoor environment can be modeled by planar patches).

6 Conclusion

We have presented a scene reconstruction algorithm which uses multiple omni-directional cylindrical panoramic images obtained from different non co-linear viewpoints. In our approach image correspondences are established by projecting points from scene space to different views and evaluating a similarity measure between local windows centered at the projected points. The main advantage of this scene space-based approach is that range maps acquired from different pairs of images are easily fused into a more robust and reliable range map. Our experimental results clearly demonstrate the potential of our approach. We foresee and have discussed several improvements and enhancements of the current system. Our goal is to warp panoramic images to images obtained at novel viewpoints [10, 7] and to use this ability in the context robot map learning in a manner similar to [3, 8].

References

1. A. Blake and M. Isard. The CONDENSATION algorithm — conditional density propagation and applications to visual tracking. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 361. The MIT Press, 1997.
2. Roland Bunschoten and Ben Kröse. Range estimation from a pair of omnidirectional images. To be presented at ICRA 2001 (accepted), 2000.
3. J.L. Crowley, F Wallner, and Sciele B. Position estimation using principal components of range data. In *Proceedings 1998 IEEE International Conference on Robotics and Automation*, pages 3121–3128, 1998.
4. Joris Van Dam. *Environment Modelling for Mobile Robots: Neural Learning for Sensor Fusion*. PhD thesis, Dept. of Computer Systems, University of Amsterdam, 1998.
5. Dieter Fox, Wolfram Burgard, Frank Dellaert, and Sebastian Thrun. Monte carlo localization: Efficient position estimation for mobile robots. In *Proceedings of the 6th National Conference on Artificial Intelligence (AAAI-99); Proceedings of the 11th Conference on Innovative Applications of Artificial Intelligence*, pages 343–349, Menlo Park, Cal., July 18–22 1999. AAAI/MIT Press.
6. J. Heikkilä and O. Silvén. Calibration procedure for short focal length off-the-shelf CCD cameras. In *Proceedings of the 13th International Conference on Pattern Recognition*, pages 166–170, Vienna, Austria, 1996.
7. Sing Bing Kang and Pavan K. Desikan. Virtual navigation of complex scenes using clusters of cylindrical panoramic images. Technical Report CRL 97/5, Cambridge Research Laboratory, Digital Equipment Corporation, 1997.
8. B.J.A. Kröse. An efficient representation of the robot’s environment. In *Proceedings Intelligent Autonomous Systems*, pages 589–595, Venice, Italy, 2000. IOS press.
9. B.J.A Kröse, N. Vlassis, R Bunschoten, and Y. Motomura. A probabilistic model for appearance-based robot localization. *Image and Vision Computing Journal (in press)*, 2000.
10. Leonard McMillan and Gary Bishop. Plenoptic modeling: An image-based rendering system. In *Proceedings of SIGGRAPH 95*, pages 39–46, Los Angeles, CA, 1995.
11. Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, June 1994.
12. Tomáš Svoboda, Tomáš Pajdla, and Václav Hlaváč. Motion estimation using central panoramic cameras. In *IEEE Conference on Intelligent Vehicles*, pages 335–340, Stuttgart, Germany, 1998.
13. P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *Int Journal of Computer Vision*, 24(3):271–300, 1997.