Statistics of Visual and Partial Depth Data for Mobile Robot Environment Modeling

Luz A. Torres-Méndez1, Gregory Dudek[2]

[1] CINVESTAV Unidad Saltillo, Ramos Arizpe, Coahuila, C.P. 25900, Mexico.
[2] Centre for Intelligent Machines, McGill University, Montreal,Quebec, H3A 2A7, CA
abril.torres@cinvestav.edu.mx,dudek@cim.mcgill.ca

Abstract. In mobile robotics, the inference of the 3D layout of large-scale indoor environments is a critical problem for achieving exploration and navigation tasks. This article presents a framework for building a 3D model of an indoor environment from partial data using a mobile robot. The modeling of a large-scale environment involves the acquisition of a huge amount of range data to extract the geometry of the scene. This task is physically demanding and time consuming for many real systems. Our approach overcomes this problem by allowing a robot to rapidly collect a set of intensity images and a small amount of range information. The method integrates and analyzes the statistical relationships between the visual data and the limited available depth on terms of small patches and is capable of recovering complete dense range maps. Experiments on real-world data are given to illustrate the suitability of our approach.

1 Introduction

One of the major goals of mobile robot research is the creation of a 3D model from local sensor data collected as the robot moves in an unknown environment. Having a mobile robot able to build a 3D map of the environment is particularly appealing as it can be used for several important applications (e.g. virtual exploration of remote locations, automatic rescue and inspection of hazardous or inhospitable environments, museums' tours, etc.). All these applications depend on the transmission of meaningful visual and geometric information. To this end, suitable sensors to densely cover the environment are required. Since all sensors are imperfect, sensor inputs must be used in a way that enables the robot to interact with its environment successfully in spite of measurement uncertainty. One way to cope with the accumulation of uncertainty is through *sensor fusion*, as different types of sensors can have their data correlated appropriately, strengthening the confidence of the resulting percepts well beyond that of any individual sensor's readings.

A typical 3D model acquisition pipeline is composed by a 3D scanner to acquire precise geometry, and a digital camera to capture appearance information. Photometric details can be acquired easily, however, to acquire dense range maps is a time and energy consuming process, unless costly and/or sophisticated hardware is used. Thus, when building 3D models or map representations of large

scenes, is desirable to simplify the way range sensor data is acquired so that time and energy consumption can be minimized. This can be achieved by acquiring only partial, but reliable, depth information.

Surface depth recovery is essential in multiple applications involving robotics and computer vision. In particular, we investigate the autonomous integration of incomplete sensory data to build a 3D model of an unknown large-scale [1] indoor environment. Thus, the challenge becomes one of trying to extract, from the sparse sensory data, an overall concept of shape and size of the structures within the environment.

We explore and analyze the statistical relationships between intensity and range data in terms of small image patches. Our goal is to demonstrate that the surround (context) statistics on both the intensity and range image patches can provide information to infer the complete 3D layout of space. It has been shown by Lee *et al.* [6] that although there are clear differences between optical and range images, they do have similar second-order statistics and scaling properties (i.e., they both have similar structure when viewed as random variables). Our motivation is to exploit this fact and also that both video imaging and *limited* range sensing are ubiquitous readily-available technologies while complete volume scanning is prohibitive on most mobile platforms.

In summary, this research answers the question of how the statistical nature of visual context can provide information about its geometric properties. In particular, how can the statistical relationships between intensity and range data be modeled reliably such that the inference of unknown range be as accurate as possible?

2 Related Work

Most prior work focuses on the extraction of geometric relationships and calibration parameters in order to achieve realistic and accurate representations of the world. In most cases it is not easy to extract the required features, and human intervention is often required. Moreover, real world environments include a large number of characteristics and properties due to scene illumination, sensor geometry, object geometry, and object reflectance, that have to be taken into account if we want to have a realistic and robust representation.

Dense stereo vision gained popularity in the early 1990's due to the large amount of range data that it could provide [8]. In mobile robotics, a common setup is the use of one or two cameras mounted on the robot to acquire depth information as the robot moves through the environment [9]. The cameras must be precisely calibrated for reasonably accurate results. The depth maps generated by stereo under normal scene conditions (i.e., no special textures or structured lighting) suffer from problems inherent in window-based correlation. These problems manifest as imprecisely localized surfaces in 3D space and as hallucinated surfaces that in fact do not exist. Other works have attempted to model 3D

[1] Large-scale space is defined as a physical space that cannot be entirely perceived from a single vantage point [5].

objects from image sequences [2, 11], with the effort of reducing the amount of calibration and avoiding restriction on the camera motion. In general, these methods derive the epipolar geometry and the trifocal tensor from point correspondences. However, they assume that it is possible to run an interest operator such as a corner detector to extract from one of the images a sufficiently large number of points that can then be reliably matched in the other images. It appears that if one uses information of only one type, the reconstruction task becomes very difficult and works well only under narrow constraints.

There is a vast body of research work using laser rangefinders for different applications, particularly, in the 3D reconstruction problem [10, 12]. However, the limitations of using only one type of sensor have increased the interest in fusing two or more type of data. Specifically, the fusing of intensity and range information for 3D model building and virtual reality applications [7, 12] with promising results. These methods use dense intensity images to provide photometric detail which can be registered and fused with range data to provide geometric detail. However, there is one notable difference, in our work the amount of range data acquired is *very* small compared to the intensity data.

3 Our framework

This research work focuses on modeling man-made large-scale indoor environments. Man-made indoor environments have inherent geometric and photometric characteristics that can be exploited to help in the reconstruction. We use a robot to navigate the environment, and together with its sensors, captures the geometry and appearance of the environment in order to build a complete 3D model.
We divide the 3D environment modeling in the following stages:

- *data acquisition and registration* of the intensity and partial range data;
- *range synthesis*, which refers to the estimation of dense range maps at each robot pose;
- *data integration* of the local dense range maps to a global map; and
- *3D model representation.*

In this paper, we only cover in detail the first two stages (see [13]). Experiments were carried out into two environments of different size and type of objects they contain. The first environment is a medium-size room (9.5m $\times$ 6m $\times$ 3m. It contains the usual objects in offices and labs (e.g., chairs, tables, computers, tools, etc.) The second environment is larger (2m $\times$ 20m $\times$ 3m) and corresponds to the corridors of our building. This environment is mostly composed of walls, doors, windows. The results are shown in each of stage described next.

4 Data Acquisition and Registration

The main aspect of our data acquisition system relies on *how* the data is acquired, which provides two important benefits: *i*) it allows the robot to rapidly collect

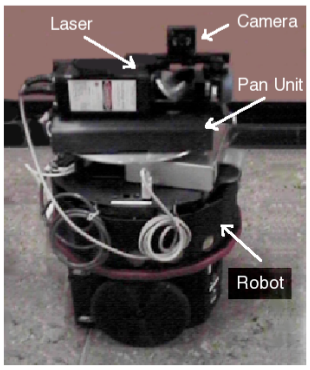


Fig. 1. Our mobile robot with the 2D laser range finder and camera mounted on it.

sparse range data and intensity images while navigating the environment to be modeled, and *ii*) it facilitates the sensor-to-sensor registration. The first benefit is essential when dealing with large environments, where the acquisition of huge amount of range data is a time consuming and impractical task. The second benefit is related to the complexity of registering different types of sensor data, which have different projections, resolutions and scaling properties. To this end, an image-based technique is presented for registering the range and intensity data that takes advantage of the way data is acquired.

The mobile robot used in our experiments is a Nomad Super Scout II, manufactured by Nomadics, Inc., retrofitted and customized for this work. On top of the robot we have assembled a system consisting of a 2D laser rangefinder, from Accuity Research, Inc., and a CCD Dragonfly camera from Point Grey Research (see Figure 1) both mounted in a *pan unit*.

The camera is attached to the laser in such a way that their center of projections (optical center for the camera and mirror center for the laser) are aligned to the center of projection of the pan unit. This alignment facilitates the registration between the intensity and range data, as we only need to know their projection types in order to do image mapping.

We assume dense and uniformly sampled intensity images, and sparse but uniformly sampled range images. Since taking images from the camera is an effortless task, sampling of intensity images occurs more often than that of range images. The area covered by the sampling data is equal at each robot pose, it covers approximately a view of 90^o. However, the amount of range data may vary depending essentially on the sampling strategy.

4.1 Acquiring Partial Range Data

The spinning mirror (y-axis) of the laser rangefinder and panning motor (x-axis) combine to allow the laser to sweep out a longitude-latitude sphere. Since each step taken by the pan unit can be programmed, we can have different sampling

strategies to acquire sparse range data. We adopt a simple heuristic for sampling which depends on how far the robot is from the objects/walls in the scene. Thus, as the robot gets closer to objects, the subsampling can be sparser since no much details are lost, compared to when the robot is located far away.

4.2 Acquiring the Cylindrical Panorama Mosaic

A cylindrical panorama is created by projecting images taken from the same viewpoint, but with different viewing angles onto a cylindrical surface. Each scene point $\mathbf{P} = (x, y, z)^T$ is mapped to cylindrical coordinate system (ψ, v) by

$$\psi = \arctan(\frac{x}{z}), \quad v = f\frac{y}{\sqrt{x^2 + z^2}}. \tag{1}$$

where ψ is the panning angle, v is the scanline, and f is the camera's focal length. The projected images are "stitched" and correlated. The cylindrical image is built by translating each component image with respect to the previous one. Due to possible misalignments between images, both a horizontal t_x and a vertical t_y translations are estimated for each input image. We then estimate the incremental translation $\delta\mathbf{t} = (\delta t_x, \delta t_y)$ by minimizing the intensity error between two images,

$$E(\delta\mathbf{t}) = \sum_{\mathbf{i}} [\mathbf{I_1}(\mathbf{x'_i} + \delta\mathbf{t}) - \mathbf{I_0}(\mathbf{x_i})]^\mathbf{2}, \tag{2}$$

where $\mathbf{x}_i = (x_i, y_i)$ and $\mathbf{x}'_i = (x'_i, y'_i) = (x_i + t_x, y_i + t_y)$ are corresponding points in the two images, and $\mathbf{t} = (\mathbf{t_x}, \mathbf{t_y})$ is the global translational motion field which is the same for all pixels. After a first order Taylor series expansion, the above equation becomes

$$E(\delta\mathbf{t}) \approx \sum_{\mathbf{i}} [\mathbf{g_i^T}\delta\mathbf{t} + \mathbf{e_i}]^\mathbf{2}, \tag{3}$$

where $e_i = I_1(x'_i) - I_0(x_i)$ is the current intensity or color error, and $g_i^T = \nabla I_1(x'_i)$ is the image gradient of I_1 at x'_i. This minimization problem has a simple least-squares solution,

$$(\sum_i g_i g_i^T)\delta\mathbf{t} = -(\sum_{\mathbf{i}} [\mathbf{e_i}\mathbf{g_i}]). \tag{4}$$

The complexity of the registration lies on the amount of overlap between the images to be aligned. In our experimental apparatus, as the panning angles at which images are taken is known, the overlap can be as small as 10% and still be able to align the images. To reduce discontinuities in intensity between images, we weight each pixel in every image proportionally to their distance to the edge of the image (i.e., it varies linearly from 1 at the centre of the image to 0 at the edge), so that intensities in the overlap area show a smooth transition between intensities in one image to intensities of the other image. A natural weighting function is the *hat function*,

$$\mathbf{w}(x, y) = \|\frac{h/2 - x}{h/2}\| - \|\frac{w/2 - y}{w/2} \tag{5}$$

Fig. 2. A cylindrical panorama.

where h and w are the height and the width of the image. In our experiments, the pan unit rotates at every 18 degrees. Figure 2 presents a 180^o cylindrical panorama constructed using the technique described above.

4.3 Camera-Laser Data Registration: Panorama with depth

The panoramic image mosaic and the incomplete spherical range data must be registered for the range synthesis. An image-based technique, similar to that in [1], is used that recovers the projective model transformation by computing a direct mapping between the points in the data sets. First, we need to convert the spherical range image to a cylindrical representation similar to that of the panoramic image mosaic, to do that the radius of the cylindrical range image must be equal to the camera's focal length. This mapping is given by

$$\mathbf{P}(r,\theta,\phi) \mapsto \mathbf{P}(r,\phi,\frac{f}{\tan\theta}) \mapsto \mathbf{P}(r,\phi,h) \tag{6}$$

where r represents the distance from the center of the cylinder to the point, h is the height of the point projected on the cylinder, ϕ is the azimuth angle and f the focal length of the camera (see Fig. 3). Again, this data is sampled on a cylindrical grid (ϕ, h) and represented as a cylindrical image.

Once having the intensity and range data in similar cylindrical image representations, a global mapping between them is computed. For a point $x_l(\phi, h)$ in the cylindrical laser image, its corresponding point in the panoramic mosaic $x_c(u, v)$ is

$$\begin{aligned} u &= a\phi + \alpha, \\ v &= f\frac{Y - \Delta Y}{r} = f\frac{Y}{r} - f\frac{\Delta Y}{r} \;\; = \;\; bh - f\frac{\Delta Y}{r} \end{aligned} \tag{7}$$

where a and b are two warp parameters that will account for difference in resolution between the two images, α aligns the pan rotation, ΔY is a vertical translation between the sensors, and $Y = rh/\sqrt{f^2 + h^2}$ is the height of the 3D point $X(r, \phi, h)$. Since f, ΔY, and the r remain fixed through the experimental setup, the term $f\frac{\Delta Y}{r}$ can be approximated to a constant β. Thus, the general warp equations are:

$$u = a\phi + \alpha, \quad v = bh + \beta \tag{8}$$

The warp parameters (a, b, α, β) are computed by minimizing the sum of the squared error of two or more corresponding points in the two images. The initial

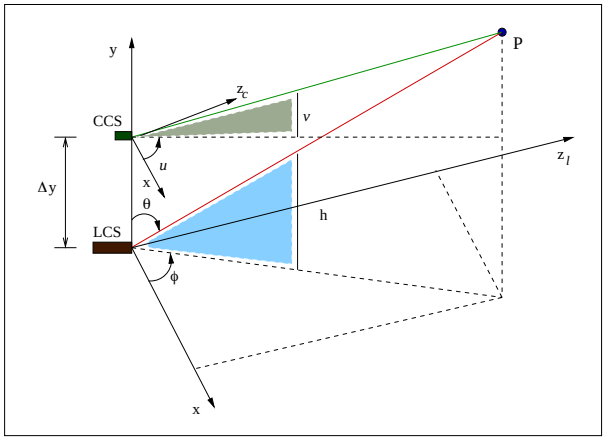


Fig. 3. Projection of the 3D point P onto cylindrical coordinates: (ϕ, h) for the range data and (u, v) for the panoramic mosaic.

estimate places the panorama mosaic nearly aligned with the range data, with a moderate translation or misalignment typically of about 5 to 7 pixels. To correct this, a *local alignment* is performed using the set of corresponding control points.

For the arrangement used in these experiments, $f = 300$ pixels, $\Delta Y = 5$ cm and the range of the points is $r = 5 - 8$ m, and β is between 6 to 10 pixel units. Figure 4 shows a samples of the registered panorama mosaic (top) and range image (bottom). It is important to note that the registration was computed using only partial range data as an input, but we show the complete range map for viewing purposes.

Fig. 4. A registered intensity (top) and range (bottom) data collected from our lab.

5 Range synthesis

After registering the intensity and partial range data at every robot pose, we apply our range synthesis method. The following sections detail our statistical learning method for depth recovery. Specifically, we estimate dense or high resolution range maps of indoor environments using only intensity images and sparse partial depth information. Markov Random Field (MRF) models are proposed as a viable stochastic model for the spatial distribution of intensity and range data. This model is trained using the (local) relationships between the observed range data and the variations in the intensity images and then used to compute unknown depth values. The MAP-MRF estimation is achieved by using the belief propagation (BP) algorithm.

5.1 The MRF Model

The range estimation problem can be posed as a labeling problem. A labeling is specified in terms of a set of *sites* and a set of *labels*. In our case, sites represent the pixel intensities in the matrix I and the labels represent the depth values in R. Let $\mathcal{S}$ index a discrete set of M sites $\mathcal{S} = \{s_1, s_2, ..., s_M\}$, and $\mathcal{L}$ be the set of corresponding labels $\mathrm{L} = \{l_1, l_2, ..., l_M\}$, where each l_i takes a depth value. The inter-relationship between sites and labels define the *neighborhood system* $\mathcal{N} = \{N_s \mid \forall s \in \mathcal{S}\}$, where N_s is the set of *neighbors* of s, such that (1) $s \notin N_s$, and (2) $s \in N_r \iff r \in N_s$. Each site s_i is associated with a random variable (*r.v.*) F_i. Formally, let $\mathrm{F} = \{F_1, ..., F_M\}$ be a random field defined on $\mathcal{S}$, in which a r.v. F_i takes a value f_i in $\mathcal{L}$. A realization $f = f_1, ..., f_M$, is called a *configuration* of F, corresponding to a realization of the field. The r.v. F defined on $\mathcal{S}$ are related to one another via the neighborhood system $\mathcal{N}$. F is said to be an MRF on $\mathcal{S}$ with respect to $\mathcal{N}$ iff the following two conditions are satisfied [4]:

$$1)\ P(f) > 0 \text{ (positivity)}, \quad \text{and } 2)\ P(f_i \mid f_{\mathcal{S}-\{i\}}) = P(f_i \mid f_{N_i}) \quad \text{(Markovianity)}.$$

where $\mathcal{S} - \{i\}$ is the set difference, $f_{\mathcal{S}-\{i\}}$ denotes the set of labels at the sites in $\mathcal{S} - \{i\}$ and $f_{N_i} = \{f'_i \mid i' \in N_i\}$ stands for the set of labels at the sites neighboring i. The Markovianity condition describes the local characteristics of F. The depth value (label) at a site is dependent only on the augmented voxels (containing intensity and/or range) at the neighboring sites. In other words, only neighboring augmented voxels have direct interactions on each other.

The choice of N together with the conditional probability distribution of $P(f_i \mid f_{\mathcal{S}-\{i\}})$, provides a powerful mechanism for modeling spatial continuity and other scene features. On one hand, we choose to model a neighborhood N_i as a square mask of size $n \times n$ centered at pixel location i, where only those augmented voxels with already assigned intensity and range values are considered in the synthesis process. On the other hand, calculating the conditional probabilities in an explicit form to infer the exact maximum *a posteriori* (MAP) in MRF models is intractable. We cannot efficiently represent or determine all the possible combinations between pixels with its associated neighborhoods. Various techniques exist for approximating the MAP estimate, such as Markov Chain

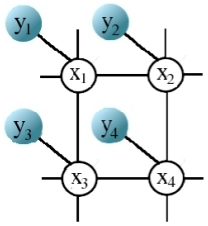


Fig. 5. Pairwise Markov network for the range estimation problem.

Monte Carlo (MCMC), iterated conditional modes (ICM), etc. We avoid the computational expense of sampling from a probability distribution and use the belief propagation algorithm to compute marginal probabilities.

5.2 MAP-MRF using Belief Propagation (BP)

In order to propagate evidence, we use a pairwise Markov network. BP efficiently estimates Bayesian beliefs in the MRF network by iteratively passing messages between neighboring nodes. The pairwise Markov network for the range estimation problem is shown in Fig. 5, where the observation node y_i is a neighborhood in intensity centered at voxel location i, and the hidden nodes x_i are the depth values to be estimated, but also hidden nodes contain the already available range data (as image patches), whose beliefs remain fixed at all times.

Learning the Compatibility Functions A *local* subset of patches containing intensity and range are used as training pairs to learn the compatibility functions. This reflects our heuristics about how the intensity values locally provide knowledge about the type of surface that intensity value belongs to.

As in [3], we use the overlapping information from the intensity image patches themselves, to estimate the compatibilities $\Psi(x_j, x_k)$ between neighbors. Let k and j be two neighboring intensity image patches. Let d^l_{jk} be a vector of pixels of the lth possible candidate for image patch x_k which lie in the overlap region with patch j. Likewise, let d^m_{kj} be the values of the pixels (in correspondence with those of d^l_{jk}) of mth candidate for patch x_j which overlap patch k. We say that image candidates x^l_k (candidate l at node k) and x^m_j are compatible with each other if the pixels in their region of overlap agree. We assume a Gaussian noise of covariance σ_i and σ_s, respectively. Then, the compatibility matrix between range nodes k and j are defined as follows:

$$\Psi(x^l_k, x^m_j) = \exp^{-|d^l_{jk} - d^m_{kl}|^2 / 2\sigma_s^2}. \tag{9}$$

The rows and columns of the compatibility matrix $\Phi(x^l_k, x^m_j)$ are indexed by l and m, the range image candidates at each node, at nodes j and k.

We say that a range image patch candidate x^l_k is compatible with an observed intensity image patch y_0 if the intensity image patch y^l_k, associated with the

range image patch candidate x_k^l in the training database matches y_0. Since it will not exactly match, we must again assume "noisy" training data and define the compatibility

$$\Phi(x_k^l, y_k) = \exp^{-|y_k^l - y_o|^2/2\sigma_s^2}. \tag{10}$$

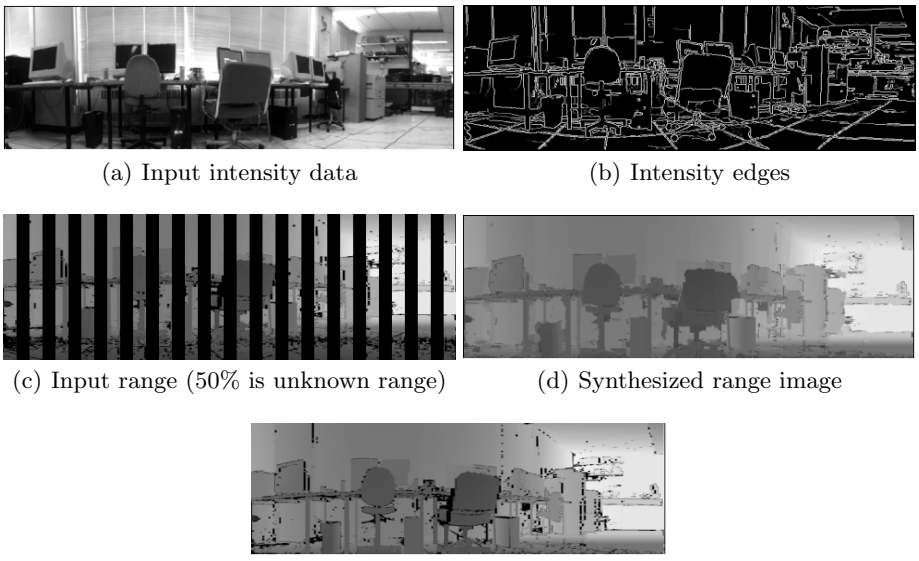
The maximum a posteriori (MAP) range image patch for node i is:

$$x_{iMAP} = \arg\max_{\mathbf{x}_i} \Phi(x_i, y_i) \prod_{j \in \mathrm{N}(i)} M_{ji}(x_i). \tag{11}$$

where $N(i)$ are all node neighbors of node i, and M_{ji} is the message from node j to node i and is computed as follows (Z is the normalization constant):

$$M_{ij}(x_j) = Z \sum_{x_i} \Psi(x_i, x_j) \Phi(x_i, y_i) \prod_{k \in \mathrm{N}(i) \setminus \{j\}} M_{ki}(x_i) \tag{12}$$

An example of applying our range synthesis algorithm is shown in Fig. 6. In (a) is the input intensity, (b) the intensity edges and (c) the input partial range data, where 50% of the total range is unknown. The resulted synthesized range image is shown in (d), and the ground truth range image in (e), for comparison purposes. The MAR error for this example is 7.85 cm.



(a) Input intensity data (b) Intensity edges

(c) Input range (50% is unknown range) (d) Synthesized range image

(e) Ground truth range

Fig. 6. Results on dense range map estimation. (a)-(c) Input data to our range synthesis algorithm. (b) The synthesized range image and (e) the ground truth range.

6 Conclusions

The ability to reconstruct a 3D model of an object or scene greatly depends on the type, quality and amount of information available. The data acquisition framework described here was designed to speed up the acquisition of range data by obtaining a relatively small amount of range information from the scene to be modeled. By doing so, we compromise the accuracy of our final representation. However, since we are dealing with man-made environments, the coherence of surfaces and their causal inter-relationships with the photometric information facilitate the estimation of complete range maps from the partial range data.

7 Acknowledgements

We would like to thank to Conacyt and NSERC for funding this research work.

References

1. D. Cobzas. *Image-Based Models with Applications in Robot Navigation.* PhD thesis, University of Alberta, Canada, 2003.
2. A.W. Fitzgibbon and A. Zisserman. Automatic 3d model acquisition and generation of new images from video sequences. In *Proceedings of European Signal Processing Conference*, pages 1261–1269, 1998.
3. W.T. Freeman, E.C. Pasztor, and O.T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 20(1):25–47, 2000.
4. J.M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. In *Unpublished*, 1971.
5. B. Kuipers. Modelling spatial knowledge. *Cognitive Science*, 2:1291–153, 1978.
6. A. Lee, K. Pedersen, and D. Mumford. The complex statistics of high-contrast patches in natural images, 2001. private correspondence.
7. M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, J. Ginsberg, J. Shade, and D. Fulk. The digital michelangelo project: 3d scanning of large statues. In *SIGGRAPH*, July 2000.
8. J.J. Little and W.E. Gillett. Direct evidence for oclussion in stereo and motion. *Image and Vision Computing*, 8:328–340, November 1990.
9. D. Murray and J. J. Little. Using real-time stereo vision for mobile robot navigation. *Autonomous Robots*, 8(2):161–171, 2000.
10. L. Nyland, D. McAllister, V. Popescu, C. McCue, A. Lastra, P. Rademacher, M. Oliveira, G. Bishop, G. Meenakshisundaram, M. Cutts, and H. Fuchs. The impact of dense range data on computer graphics. In *Proceedings of Multi-View Modeling and Analysis Workshop (MVIEW part of CVPR)*, page 8, June 1999.
11. M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool. Metric 3d surface reconstruction from uncalibrated images sequences. In *Proceedings of SMILE Workshop (post-ECCV)*, pages 138–153, 1998.
12. I. Stamos and P.K. Allen. 3d model construction using range and image data. In *CVPR*, June 2000.
13. L. Abril Torres-Méndez. *Statistics of Visual and Partial Depth Data for Mobile Robot Environment Modeling.* PhD thesis, McGill University, 2005.