Range synthesis for 3D Environment Modeling

Luz A. Torres-Méndez and Gregory Dudek
Center for Intelligent Machines, McGill University
Montreal, Quebec H3A 2A7, CA
{latorres,dudek}@cim.mcgill.ca

***Abstract*— This paper examines a novel method we have developed for computing range data in the context of mobile robotics. Our objective is to compute dense range maps of locations in the environment, but to do this using intensity images and very limited range data as input. We develop a statistical learning method for inferring and extrapolating range data from a combination of a single video intensity image and a limited amount of input range data. Our methodology is to compute the relationship between the observed range data and the variations in the intensity image, and use this to extrapolate new range values. These variations can be efficiently captured by the neighborhood system of a Markov Random Field (MRF) without making any strong assumptions about the kind of surfaces in the world. Experimental results show the feasibility of our method.**

I. INTRODUCTION

Mobile robots commonly navigate using range data, be it from a precomputed map, a laser range scanner or from stereo. A particularly common simplifying assumption is that the world is essentially two-dimensional, and that a 2D ”slice” through the 3D world is sufficient to represent 3D structure of interest. This assumption is used with such prevalence in part because the acquisition of complete range maps (i.e. volume scans) is is costly and complicated. In this paper, we bypass this assumption and will present an approach to the acquisition of estimates 3D data from a combination of a single video intensity image and a limited amount of input range data.

The motivation is to exploit the fact that both video imaging and *limited* range sensing are ubiquitous readily-available technologies while complete volume scanning is prohibitive on most mobile platforms. Our approach is based on the observation that we can use the video imagery we collect to extrapolate or interpolate a limited amount of (sparse) range data to compute a dense map. Note that we are not interested in simply inferring a few missing pixels, but in synthesizing a complete range map from as little as a few scans of a laser line striping device across the environment.

Our approach to this problem is to solve the range data inference problem as an extrapolation problem. This, in turn, is cast in statistical terms by approximating the *composite* of range and intensity at each point as a Markov process. Range data inference is accomplished by using the statistics of the observed range data to determine the behavior of the Markov process and solve for the unknown range data. Critical to the processes is the presence of intensity data at each pixel where range is being inferred. Intuitively, this intensity data provides at least to kinds of information: knowledge of when the surface being synthesized is smooth, and knowledge of when there is a high probability of a variation in depth. On strength of our approach, however, is that we learn these inferences from the observed data, and do not need to fabricate or hypothesize constraints that might be inapplicable to a particular environment.

In the present paper we use ground-truth data from Oak Ridge National Labs. As such, while our target application is mobile robotics, this paper does not explicitly address the issues of navigation and data acquisition.

Our approach is based on the assumption that the pixels constituting both the range and intensity images acquired in an environment, can be regarded as the results of pseudo-random processes, but that these random processes exhibit useful structure. In particular, we exploit the assumption that range and intensity images are correlated, albeit in potentially complicated ways. Secondly, we assume that the variations of pixels in the range and intensity images are related to the values elsewhere in the image(s) and that these variations can be efficiently captured by the neighborhood system of a Markov Random Field. Both these assumptions have been considered before [6], [3], [17], [2], [7], but we they have never been exploited in tandem.

In this paper we briefly consider some of the related prior work, outline our formalism and the algorithm we use to infer range data, and then present several types of experimental data showing the properties of this approach.

II. BACKGROUND

The inference of 3D models of a scene is a problem that subsumes a large part of computer vision research over the last 30 years. In the context of this paper we will consider only a few representative solutions.

Over the last decade laser rangefinders have become affordable and available but their application to building full 3D environment models, even from a single viewpoint, remains costly or difficult in practice. In particular, while laser line scanners based on either triangulation and/or time-of-flight are ubiquitous, full volume scanners tend to be much more complicated and physically sensitive. As a result, the acquisition of *dense, complete* 3D range maps is still a pragmatic challenge even if the availability of laser range scanners is presupposed.

Much of the previous work on environment modeling uses one of either photometric data or geometric data [1], [8], [5], [12] to reconstruct a 3D model of an scene. For example, Fitzgibbon and Zisserman [5] proposed a method that sequentially retrieves the projective calibration of a complete image sequence based on tracking corner and/or line features over two or more images, and reconstructs each feature independently in 3D. Their method solves the feature correspondence problem based on the fundamental matrix and tri-focal tensor, which encode precisely the geometric constraints available from two or more images of the same scene from different viewpoints. Related work includes that of Pollefeys et. al. [12]; they obtain a 3D model of an scene from image sequences acquired from a freely moving camera. The camera motion and its settings are unknown and there is no prior knowledge about the scene. Their method is based on a combination of the projective reconstruction, self calibration and dense depth estimation techniques. 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.

Shape-from-shading is related in spirit to what we are doing, but is based on a rather different set of assumptions and methodologies. Such method [9], [11] reconstruct a 3D scene by inferring depth from a 2D image; in general, this task is difficult, requiring strong assumptions regarding surface smoothness and surface reflectance properties. Recent work has considered the use of both intensity data as well as range measurements. Several authors [13], [4], [14], [10], [15] have obtained promising results. Pulli et al. [13] address the problem of surface reconstruction by measuring both color and geometry of real objects and displaying realistic images of objects from arbitrary viewpoints. They use a stereo camera system with active lighting to obtain range and intensity images as visible from one point of view. The integration of the range data into a surface model is done by using a robust hierarchical space carving method. The integration of intensity data with range data has been proposed [14] to help define the boundaries of surfaces extracted from the 3D data, and then a set of heuristics are used to decide what surfaces should be joined. For this application, it becomes necessary to develop algorithms that can hypothesize the existence of surface continuity and intersections among surfaces, and the formation of composite features from the surfaces. However, one of the main issues in using the above configurations is that the acquisition process is very expensive because dense and complete intensity and range data are needed in order to obtain a good 3D model. As far as we know, there is no method that bases its reconstruction process on having a small amount of intensity and/or range data and synthetically estimating the areas of missing information by using the current available data. In particular, such a method is feasible in man-made environments, which, in general, have inherent geometric constraints, such as planar surfaces.

III. THE ALGORITHM

As noted above our objective is to compute range values where only intensity is known. We will do this by incrementally computing a single range value at a time by using neighboring locations where both range and intensity is available. We assume that the intensity and range data is already registered [1].

We use Markov Random Fields (MRF) as a model that captures characteristics of the relationship between intensity and range data in a neighborhood of a given voxel, i.e. the data in a voxel are determined by its immediate neighbors (and prior knowledge) and not on more distant voxels (the locality property). While this assumption is not strictly valid, our results seem very satisfactory; the implications of this are discussed later. The other property that we exploit is limited stationarity, i.e. different regions of an image are always perceived to be similar. This property is true for textures but not for more general classes of images representing scenes containing one or more objects. In our algorithm, we synthesize a depth value so that it is locally similar to some region not very far from its location. The process is completely deterministic, meaning that no explicit probability distribution needs to be constructed.

A. Synthesizing range

We focus on our development of a set of **augmented voxels V** that contain intensity and range information (where the range is initially unknown for some of them). Thus, $\mathbf{V} = (\mathrm{I}, \mathrm{R})$, where I is the matrix of known pixel intensities and R denotes the matrix of incomplete pixel depths. We are interested only in a set of such augmented voxels such that one voxel lies on each ray that intersects each pixel of the input image I, thus giving us a registered range image R and intensity image I.

Let $Z_m = (x,y) : 1 \leq x,y \leq m$ denote the $m \times m$ integer lattice (over which the images are described); then $\mathrm{I} = \{I_{x,y}\}$, $(x,y) \in Z_m$, denotes the gray levels of the input image, and $\mathrm{R} = \{R_{x,y}\}$, $(x,y) \in Z_m$ denotes the depth values. We model $\mathbf{V}$ as an MRF. Thus,

[1] In practice this registration could be computed as a first step, but we omit this in the current presentation.

we regard I and R as a random variables. For example, $\{\mathrm{R} = r\}$ stands for $\{R_{x,y} = r_{x,y}, (x,y) \in Z_m\}$. Given a *neighborhood system* $\mathscr{N} = \{\mathscr{N}_{x,y} \in Z_m\}$, where $\mathscr{N}_{x,y} \subset Z_m$ denotes the neighbors of (x,y), such that, (1) $(x,y) \notin \mathscr{N}_{x,y}$, and (2) $(x,y) \in \mathscr{N}_{k,l} \Longleftrightarrow (k,l) \in \mathscr{N}_{x,y}$. An MRF over $(Z_m, \mathscr{N})$ is a stochastic process indexed by Z_m for which, for every (x,y) and every $v = (i,r)$ (i.e. each augmented voxel depends only on its immediate neighbors),

$$\begin{aligned} &P(V_{x,y} = v_{x,y} \,|\, V_{k,l} = v_{k,l}, (k,l) \neq (x,y)) \\ &\quad = P(V_{x,y} = v_{x,y} \,|\, V_{k,l} = v_{k,l}, (k,l) \in \mathscr{N}_{x,y}), \end{aligned} \tag{1}$$

The choice of $\mathscr{N}$ together with the conditional probability distribution of $P(\mathrm{I} = i)$ and $P(\mathrm{R} = r)$, provides a powerful mechanism for modeling spatial continuity and other scene features. On one hand, we choose to model a neighborhood $\mathscr{N}_{x,y}$ as a square mask of size $n \times n$ centered at the voxel location (x,y). This neighborhood is causal, meaning that only those voxels already containing both, intensity and range information are considered for the synthesis process. On the other hand, calculating the conditional probabilities in an explicit form is an infeasible task since we cannot efficiently represent or determine all the possible combinations between augmented voxels with its associated neighborhoods. Therefore, we avoid the usual computational expense of sampling from a probability distribution (Gibbs sampling, for example), and synthesize a depth value $R_{x,y}$ deterministically by selecting the range value $R_{k,l}$ from the augmented voxel whose neighborhood most resembles the region being filled in, i.e.,

$$V_{best} = \underset{(k,l) \in \mathscr{A}}{\mathbf{argmin}} \; \| V_{x,y} - V_{k,l} \|, \tag{2}$$

where $\mathscr{A} = \{\mathscr{A}_{k,l} \subset \mathscr{N}\}$ is the set of local neighborhoods, such that $1 \leq \sqrt{(k-x)^2 + (l-y)^2)} \leq d$. For each successive augmented voxel this approximates the maximum a posteriori estimate; $R(k,l)$ is then used to specify $R(x,y)$. The similarity measure $\| \,.\, \|$ is described over the partial data about locations (x,y) and (k,l) and is calculated as follows,

$$\sum_{\vec{v} \in \mathscr{N}^*} G(\sigma, \vec{v} - \vec{v}_0)[(I_{\vec{v}} - I'_{\vec{v}})^2 + (R_{\vec{v}} - R'_{\vec{v}})^2], \tag{3}$$

where $\vec{v}_0$ is the augmented voxel located at the center of the neighborhood $\mathscr{N}^*$, $\vec{v}$ is a neighboring voxel of $\vec{v}_0$. I and R are the intensity and range values of the neighboring augmented voxels of the depth value $R_{x,y} \in \vec{v}_0$ to synthesize, and I' and R' are the intensity and range values to be compared with and in which, the center voxel $\vec{v}_0$ has already assigned a depth value. G is a 2-D Gaussian kernel that gives more weight to those voxels near the center than those at the edge of the window.

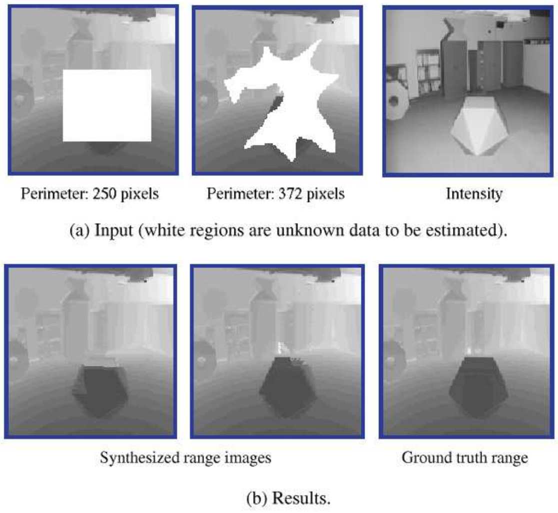
In our algorithm we synthesize one depth value at a time. In our experiments, depth values are assigned in a spiral-scan ordering, either growing inwards or outwards, depending on the shape of the area to synthesize.

IV. Experimental Results

We have tested our algorithm using data acquired in a real-world environment. The real intensity (reflectance) and range images of indoor scenes were acquired by an Odetics laser range finder mounted on a mobile platform. Images are 128×128 pixels and encompass a $60^o \times 60^o$ field of view. We start with the complete range data set as ground truth and then hold back most of the data to simulate the sparse sample of a real scanner and to provide input to our algorithm. This allow us to compare the quality of our reconstruction with what is actually in the scene. In the following we will consider various strategies for subsampling the range data.

A. Arbitrary shape of unknown range data

The first type of experiment involves the range synthesis when the unknown range data is of arbitrary shape. Particularly, we want to show how the shape influences the range estimation. In Fig. 1a, two input range images (the left and middle images) and input intensity are given. The number of pixels in each unknown area (shown in white) of both range images is 3864. The perimeters, however are different, 250 and 372 pixels, respectively. Fig. 1b shows the synthesized range images (left and middle) and the ground truth range image for comparison purposes. The residual average errors are 7.17 and 3.99, respectively. It can be seen that when synthesizing big areas of unknown range data, our algorithm performs better if the area is not compact, since combinations of already known range and



Perimeter: 250 pixels

Perimeter: 372 pixels

Intensity

(a) Input (white regions are unknown data to be estimated).

Synthesized range images

Ground truth range

(b) Results.

Fig. 1. Results on two different shapes of unknown range with same area 3864 pixels.

intensity give more information about the geometry of the scene.

B. Limited dense range

We now show some experiments where the initial range is a window of size $p \times q$ and at position (r_x, r_y) on the intensity image. Fig. 2a shows the intensity image (left) of size 128×128 and the initial range (right), a window of size 64×64, i.e. only the 25% of the total range is known. The size of the neighborhood is 5×5 pixels. The synthesized range data obtained after running our algorithm is shown in the left side of Fig. 2b; for purposes of comparison, we show the complete real range data (right side). It can be seen that the synthesized range is very similar to the real range. The Odetics LRF uses perspective projection, so the image coordinate system is spherical. To calculate the residual errors, we first convert the range images to the Cartesian coordinate system (range units) by using the equations in [16]. For this example, the average residual error is 7.98.

C. Sparse range measurements

In a third type of experiment, the initial range data is a set of stripes with variable width along the $x-$ and $y-$axis of the intensity image. We tested with the same intensity image used in the previous section in order to compare both results. Two experiments are shown in Fig. 3. The initial range images are shown in the left column, and to their right are synthesized results. In Fig. 3a, the width of the stripes s_w, is 5 pixels, and the area with missing range data $(x_w \times x_w)$ is 25×25, i.e., 39% of the range image is known. For Fig. 3b, the values are $s_w = 3$, $x_w = 28$, in this case, only 23% of the total range is known. The average residual error (in range units) for the reconstruction are 2.37 and 3.07, respectively. In Fig. 4 a graph of the density of pixels at different depth values (scale from 0 to 255) of the original and synthesized range of Fig. 3a. Fig. 5 displays two different views using the real range and the synthesized range results of Fig. 3.

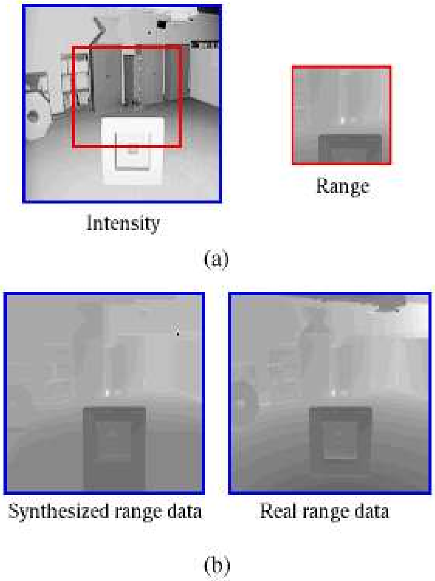
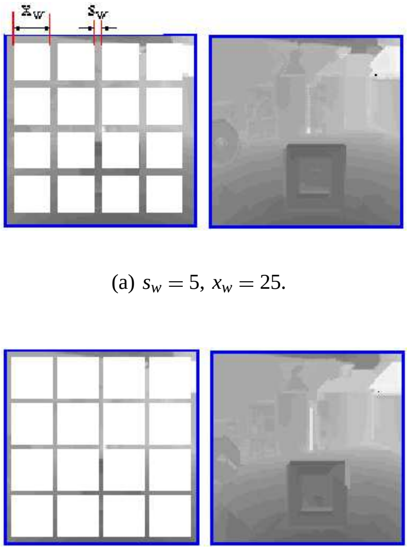


Fig. 2. Results on real data. (a) Input. (b) Results comparing synthesized range data to ground truth.



(a) $s_w = 5$, $x_w = 25$.

(b) $s_w = 3$, $x_w = 28$.

Fig. 3. Results on real data. The left column shows the initial range data and to their right is the synthesized result (the white squares represent unknown data to be estimated). Since the unknowns are withheld from genuine ground truth data, we can estimate our performance.

The results are surprisingly good in both cases. Our algorithm was capable of recovering the whole range of the image. We note, however, that results of experiments using stripes are much better than those using a window as the initial range data. Intuitively, this is because the sample spans a broader distribution of range-intensity combinations than in the local window case.

Our algorithm was tested on 30 images of common scenes found in a general indoor man-made environment.

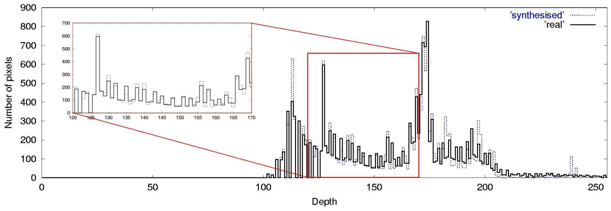


Fig. 4. Histogram of pixels at different depth values (scale 0-255) of the original and synthesized range of Fig. 3a.

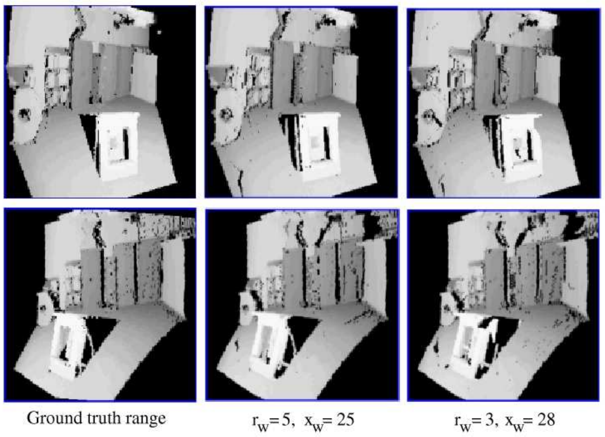
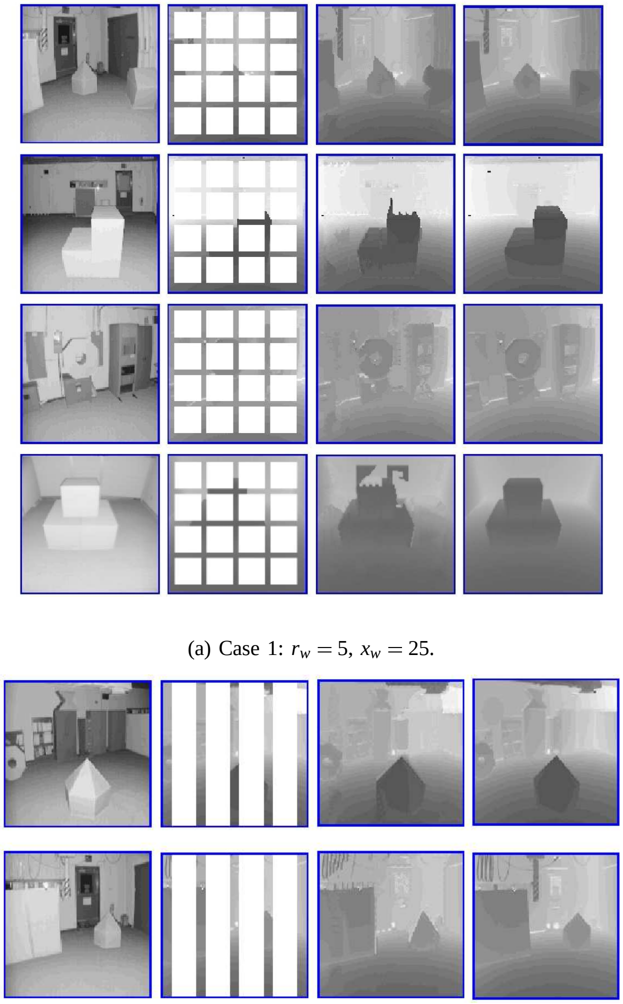


Fig. 5. Results in 3D. Two views of the real range (left column) and the synthesized results (middle and right columns) of Fig. 3.

Two cases of subsampling were used in our experiments. Case 1 is as one of the subsampling previously described, with $r_w = 5$ and $x_w = 25$, applied along $x-$ and $y-$axis. Due to space limitations, we are only showing 3 more examples of this case in Fig. 6a, the average residual errors are, from top to bottom, 2.84, 4.53, 3.32 and 5.24. For Case 2, $r_w = 8$ and $x_w = 22$, but applied only along the $x-$axis. Fig. 6b shows 2 examples of this case. Here the average residual errors are 4.17 and 5.25, respectively. The maximum average residual errors obtained from all 30 test images were 6.52 for Case I, and 11.85 for Case II.

In general, the results are good in both cases. However, sometimes our algorithm performs poorly in those locations where a high change in discontinuity exist. These changes in discontinuity can be captured by using edge information from the intensity images. Also, we can see that the order in which we choose the next depth value to synthesize will reflect the final result. With the spiral-scan ordering, there is a strong dependence from the previous assigned pixel. A better scan ordering would be to synthesize first those pixels with the maximum number of neighboring (mnn) pixels containing both, depth and intensity information. We implemented the mnn-scan ordering and also added edge information. Again our algorithm was tested on the 30 previous images. For purposes of comparison, Fig. 7 displays the results obtained for the input images shown in Fig. 6a. The average residual errors are now, from left to right, 2.68, 2.44, 3.24 and 1.75.

It is important to note, that the initial range data given as an input is crucial to the quality of the synthesis, that is, if no interesting changes exist in the range and intensity, then the task becomes difficult. However, the results presented here demonstrate that this is a viable option to facilitate environment modeling.



(a) Case 1: $r_w = 5$, $x_w = 25$.

(b) Case 2: $r_w = 8$, $x_w = 22$ along the x-axis.

Fig. 6. Examples on real data. The first and second columns are the input intensity and range data, respectively. White regions in the input data are unknown data to be inferred by the algorithm. The synthesized results are shown in the third column and, the real range images are displayed in the last column for visual comparison.

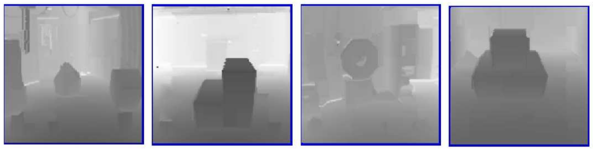


Fig. 7. Range synthesis using the mnn-scan ordering and edge information. For purposes of comparison, the input images are the same shown in Fig. 6a.

V. CONCLUSIONS AND FUTURE WORK

Knowledge of the environment is a vital component of robotics automation. However, modeling an environment is not an easy task since the acquisition of 3D geometric data is costly and complicated. In this paper we have presented an algorithm for recovering 3D geometric data given an intensity image with little associated range information.

Our approach uses Markov Random Field methods as a model that exploits the statistically observed relationship between the intensities in a neighborhood and range data to infer the unknown range. While this formalism can explicitly capture local differential geometry, we do not explicitly compute local surface properties, nor does this approach make substantive assumptions regarding surface reflectance functions of surface geometry such as smoothness. The approach does assume that the relationship between intensity and range can be expressed by a stationary distribution; an assumption that could be relaxed. While avoiding strong assumptions about the surfaces in the scene allows greater generality, it also means we do not exploit potentially useful constraint information. In ongoing work, we are examining the incorporation of more elaborate priors and geometric inferences. Also, there are a number of parameters that can greatly influence the quality of the results: the size of the neighborhood used in computing correlations, the amount of initial range and the characteristics captured in that initial range. The characterization of how these parameters effect the results is the subject of ongoing work.

VI. ACKNOWLEDGMENTS

We would like to thank the CESAR lab at Oak Ridge National Laboratory in Tennessee for making their range image database available through the website of the University of South Florida *http://marathon.csee.usf.edu/range/Database.html*.

The first author gratefully acknowledges CONACyT for providing financial support to pursue her Ph.D. studies at McGill University.

We would like to thank also the Federal Centers of Excellence (IRIS) for ongoing funding.

VII. REFERENCES

[1] P.E. Debevec, C.J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry and image-based approach. In *SIGGRAPH*, pages 11–20, 1996.

[2] A. Efros and W.T. Freeman. Image quilting for texture synthesis and transfer. In *SIGGRAPH*, pages 1033–1038, August 2001.

[3] A. Efros and T.K. Leung. Texture synthesis by non-parametric sampling. In *ICCV (2)*, pages 1033–1038, September 1999.

[4] S.F. El-Hakim. A multi-sensor approach to creating accurate virtual environments. *Journal of Photogrammetry and Remote Sensing*, 53(6):379–391, December 1998.

[5] 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.

[6] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[7] A. Hertzmann, C.E. Jacobs, N. Oliver, B. Curless, and D.H. Salesin. Images analogies. In *SIGGRAPH*, August 2001.

[8] A. Hilton. Reliable surface reconstruction from multiple range images. In *ECCV*, 1996.

[9] B.K.P. Horn and M.J. Brooks. *Shape from Shading*. MIT Press, Cambridge Mass., 1989.

[10] 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.

[11] J. Oliensis. Uniqueness in shape from shading. *Int. Journal of Computer Vision*, 6(2):75–104, 1991.

[12] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool. Automated reconstruction of 3d scenes from sequences of images. *ISPRS Journal Of Photogrammetry And Remote Sensing*, 55(4):251–267, 2000.

[13] K. Pulli, M. Cohen, T. Duchamp, H. Hoppe, J. McDonald, L. Shapiro, and W. Stuetzle. Surface modeling and display from range and color data. *Lecture Notes in Computer Science 1310*, pages 385–397, September 1997.

[14] V. Sequeira, K. Ng, E. Wolfart, J.G.M. Goncalves, and D.C. Hogg. Automated reconstruction of 3d models from real environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54:1–22, February 1999.

[15] I. Stamos and P.K. Allen. 3d model construction using range and image data. In *CVPR*, June 2000.

[16] K. Storjohann. Laser range camera modeling. Technical report, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 1990.

[17] L. Wei and M. Levoy. Fast texture synthesis using tree-structured vector quantization. In *SIGGRAPH*, pages 479–488, July 2000.