Statistics in the Image Domain for Mobile Robot 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 addresses the problem of estimating dense range maps of indoor locations using only intensity images and sparse partial depth information. Unlike shape-from-shading, we infer the relationship between intensity and range data and use it to produce a complete depth map. We extend prior work by incorporating geometric information from the available range data, specifically, we add surface normal information to reconstruct surfaces whose variations are not captured in the initial range measurements. In addition, the order on which we synthesize range values is based on a best-first approach that uses edge information from the intensity images, and isophotes lines from the available range. Our method uses Markov Random Fields to learn the statistical relationships from the available intensity and from those sparse regions where both range and intensity information is present. In contrast to classical approaches to depth recovery (i.e. stereo, shape from shading), we make only weak assumptions regarding specific surface geometries or surface reflectance functions since we compute the relationship between existing range data and the images we start with. Preliminary results on real data demonstrate that our method works reasonably well when incorporating geometric information.

1 Introduction

Surface depth recovery is one of the most classic vision problems, both because of its scientific importance and its pragmatic value. Many standard "shape-from" methods, however, are based on strong assumptions regarding scene structure or reflectance functions. While several elegant algorithms for depth recovery have been developed, the use of laser range data in many applications such as robotics has become commonplace due to their simplicity and reliability (but not their elegance, cost or physical robustness). In robotics, the fusion of range data with visual information for navigation and mapping is particularly appealing, as it can be used for several important applications. However, it is often hampered by the fact that range sensors that provide complete (2-1/2D) depth maps with a resolution akin to that of a camera, are prohibitively costly or otherwise impractical. Stereo cameras can produce volumetric scans that are economical, but they often require calibration or produce range maps that are either incomplete or of limited resolution. A particular common simplifying assumption is to represent 3D structure as a 2D "slice" through the world. However, in practice this is not sufficient to capture structures of interest.

We seek to reconstruct suitable 3D models from sparse range data sets while simultaneously facilitating the data acquisition process. It has been shown by Lee *at al.* [13] that although there are clear differences between optical and range images, they nevertheless have similar second-order statistics and scaling properties. Of course, the intimate connection between irradiance and surface normals is the basis of classical shape-from-shading. 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 remains prohibitive on most mobile platforms. It is important to highlight that we are not simply inferring a few missing pixels, but synthesizing a complete range map from as little as few laser scans across the environment.

Our methodology is to learn a statistical model of the (local) relationship between the observed range data and the variations in the intensity image and use this to compute the unknown range data. We approximate the *composite* of range and intensity at each point as a Markov process. Unknown range data is then inferred by using the statistics of the observed range data to determine the behavior of the Markov process. The presence of intensity where range data is being inferred is crucial since intensity data provides knowledge of surface smoothness and variations in depth. Our approach learns that knowledge directly from the observed data, without having to hypothesize constraints that might be inapplicable to a particular environment.

In this paper, we extend our prior work [22, 23] by in-

corporating geometric information to our framework from the available range data. More specifically, we add surface normal information to be able to reconstruct surfaces whose variations are not captured in the initial range measurements. We also have improved the order in which we recover depth values. This order plays a very important role in the quality of the results.

In the following section we consider relevant prior work. Section 3 describes our method to infer range data and the improvements to our algorithm. Section 4 tests the proposed algorithm on different configurations of experimental data. Section 5 gives details about the incorporation of surface normal information to our framework. Finally, in Section 6 we give some conclusions and future directions.

2 Previous Work

We base our range estimation process 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 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 [5,6,10,24], but they have never been exploited in tandem.

Digital inpainting [2, 3] is similar to our problem, although our domain and approach are quite different. Baker and Kanade [1] used a learned representation of pixel variation for perform resolution enhancement of face images. The processes employed to interpolate new high-resolution pixel data is similar in spirit to what we describe here, although the application and technical details differ significantly. The work by Freeman [9, 21] on learning the relationships between intrinsic images is also related.

In prior work [22, 23], we performed reconstruction by first recovering the values of those voxels for which we can make the most reliable inferences, based on the number of filled neighbors and, on the existence of probably depth discontinuities (indicated by edges on both intensity and range images). However, the extraction of edges is not always easy, and it depends on the scene (textures, changes in illumination, etc.). While the connections between intensity edges and depth discontinuities is not assured, in our approach we use it only to bias the reconstruction sequence so that spurious edges (due to albedo, for example) do not cause difficulties. In this work, we have incorporated information regarding curvilinear iso-intensity structures intensity (known as isophotes) from the available range. These isophotes, as edge information, indicate regions with depth discontinuity. Thus, as we reconstruct, we select first those voxels for reconstruction that have the largest degree of boundary constraint and that do not contain any isophotes nor any edge in its surrounding voxels.

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.

In general, the process of building a full 3D model of a real environment can be divided onto two processes: acquisition of measurements in 3D and synthesis of useful geometric models from measurements. In some cases, for example when models are generated manually, these steps may be combined. In other cases the processes of collecting sets of 3D points (often referred to as range scans), combining them onto surfaces and then generating suitable models for graphics applications entail distinct computations. In this paper we focus only on the processes of obtaining 3D data.

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 error-prone. 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.

Most of prior work on synthesis of 3D environment models uses one of either photometric data or geometric data [4,8,11,16] to reconstruct a 3D model of an scene. For example, Fitzgibbon and Zisserman [8] 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. [16]; 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 [12, 15] 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 focus on combining information from the intensity and range data for 3d model reconstruction. Several authors [7, 14, 17, 19, 20] have obtained promising results. Pulli et al. [17] 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 [19] 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.

3 Methodology

Our objective is to infer a dense range map from an intensity image and a limited amount of initial range data. At the outset, we assume that resolution of the intensity and range data is the same and that they are already registered (in practice this registration could be computed as a first step, but we omit this in the current presentation.) Note that while the process of inferring distances from intensity superficially resembles shape-from-shading, we do not depend on prior knowledge of reflectance or on surface smoothness or even on surface integrability (which is a technical precondition for most shape-from-shading methods, even where not explicitly stated).

We solve the range data inference problem as an extrapolation problem by approximating the *composite* of range and intensity at each point as a Markov process. Unknown range data is then inferred by using the statistics of the observed range data to determine the behavior of the Markov process. Critical to the processes is the presence of intensity data at each point where range is being inferred. Intuitively, this intensity data provides at least two kinds of information: (1) knowledge of when the surface is smooth, and (2) knowledge of when there is a high probability of a variation in depth. Our approach learns that information from the observed data, without having to fabricate or hypothesize constraints that might be inapplicable to a particular environment.

From previous experiments we know that reconstruction across depth discontinuities is often problematic as there is comparatively little constraint for probabilistic inference at these locations. Such locations are often identified with edges in both the range and intensity maps. In our previous experiments we used only edge information from the intensity images to lead the reconstruction process. We have noticed that our results can be improved if we also add information about linear structures from the available range data. These linear structures are called isophotes (all normals forming same angle with direction to eye). Thus, as we recover augmented voxels, we defer the reconstruction of augmented voxels close to intensity discontinuities (indicated by edges) and/or depth discontinuities (indicated by the isophotes) as much as possible.

In standard Markov Random Field methods, the assumption is that the field is updated in either stochastically or in parallel according to an iterative schedule. In practice, several authors have considered more limited update schedules. In our work, we restrict ourselves to a single update at each unknown measurement only. In our algorithm we synthesize depth value $R(x, y)$ sequentially (although this does not preclude parallel implementations). From previous experiments we know that the reconstruction sequence (the order in we choose the next depth value to synthesize) has a significant influence on the quality of the final result. For example, with the onion-peel ordering, the problem was the strong dependence from the previous assigned voxel. Our reconstruction sequence is then, to first recover the values of those augmented voxels for which we can make the most reliable inferences, based on essentially two factors: 1) the number of neighboring voxels with already assigned range and intensity and 2) the existence of intensity and/or depth discontinuities (i.e. if an edge or a linear structure exists). Priority values are computed based on these two factors and are assigned to each voxel for reconstruction, such that as we reconstruct, we select the voxel with the maximum priority value. If more than one voxel shares the same priority value, then the selection is done randomly.

3.1 The MRF model for range synthesis

Markov Random Fields (MRF) are used here as a model to synthesize range. We focus on our development of a set of **augmented voxels** $\mathbf{V}$ that contain intensity (either from grayscale or color images), edge (from the intensity image) and range information (where the range is initially unknown for some of them). Thus, $\mathbf{V} = (\mathrm{I}, \mathrm{E}, \mathrm{R})$, where I is the matrix of known pixel intensities, E is a binary matrix (1 if an edge exists and 0 otherwise) and R denotes the matrix of incomplete pixel depths. We are interested only in a set of such augmented voxels such that one augmented 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 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, 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* $\mathcal{N} = \{\mathcal{N}_{x,y} \in Z_m\}$, where $\mathcal{N}_{x,y} \subset Z_m$ denotes the neighbors of (x, y), such that, (1) $(x, y) \notin \mathcal{N}_{x,y}$, and (2) $(x, y) \in \mathcal{N}_{k,l} \Longleftrightarrow (k, l) \in \mathcal{N}_{x,y}$. An MRF over $(Z_m, \mathcal{N})$ 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),

$$P(V_{x,y} = v_{x,y} \,|\, V_{k,l} = v_{k,l}, (k, l) \neq (x, y)) = P(V_{x,y} = v_{x,y} \,|\, V_{k,l} = v_{k,l}, (k, l) \in \mathcal{N}_{x,y}), \quad (1)$$

The choice of $\mathcal{N}$ 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 $\mathcal{N}_{x,y}$ as a square mask of size $n \times n$ centered at the augmented voxel location (x, y). This neighborhood is causal, meaning that only those augmented voxels already containing information (either intensity, range or both) 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 from the augmented voxel $V_{x,y}$ with neighborhood $\mathcal{N}_{x,y}$, by selecting the range value from the augmented voxel whose neighborhood $\mathcal{N}_{k,l}$ most resembles the region being filled in, i.e.,

$$\mathcal{N}_{best} = \underset{(k,l) \in \mathcal{A}}{\textbf{argmin}} \parallel \mathcal{N}_{x,y} - \mathcal{N}_{k,l} \parallel, \quad (2)$$

where $\mathcal{A} = \{\mathcal{A}_{k,l} \subset \mathcal{N}\}$ is the set of local neighborhoods, in which the center voxel has already assigned a depth value, 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 $\parallel . \parallel$ between two generic neighborhoods $\mathcal{N}_a$ and $\mathcal{N}_b$ is defined as the weighted sum of squared differences (WSSD) over the partial data in the two neighborhoods. The "weighted" part refers to applying a 2-D Gaussian kernel to each neighborhood, such that those voxels near the center are given more weight than those at the edge of the window.

3.2 Range Synthesis Ordering

Our reconstruction sequence depends entirely on the priority values that are assigned to each augmented voxel on the boundary of the region to be synthesized. The priority computation is biased toward those voxels that are surrounded by high-confidence voxels, that are not on a isophote line, and whose neighborhood does not represent an intensity discontinuity, in other words, whose neighborhood does not have any edges on it. Furthermore, edge information is used to defer the synthesis of those voxels that are on an edge to the very end.

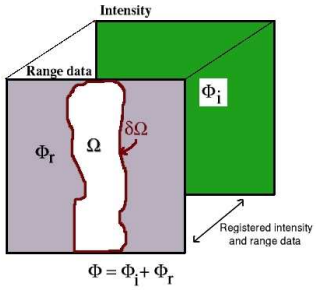

Figure 1: The notation diagram.

Figure 1 shows the basic notation used to explain how our algorithm reconstructs the unknown depth values (notation similar to that used in the inpainting literature [3]). The region to be synthesized, i.e., the *target* region is indicated by Ω, and its contour is denoted $\delta\Omega$. Only on those regions where depth discontinuities are not detected, the contour evolves inward as the algorithm progresses. The input intensity (Φ_i) and the input range (Φ_r) both together form the *source* region, and is indicated by Φ. This region Φ is used to calculate the local statistics for reconstruction. Let $V_{x,y}$ be an augmented voxel with unknown range located at the boundary $\delta\Omega$ and $\mathcal{N}_{x,y}$ be its neighborhood, which is a $n \times n$ square window centered at $V_{x,y}$. Then, for all augmented voxels $V_{x,y} \epsilon \delta\Omega$, we compute their priority value (which is going to determine the order in which they are

filled) as follows:

$$P(V_{x,y}) = C(V_{x,y})D(V_{x,y}) + 1/(1+E). \qquad (3)$$

where E is the number of edges found in $N_{x,y}$; $C(V_{x,y})$ is the *confidence* term, $D(V_{x,y})$ the *data* term. These terms are defined as follows:

$$C(V_{x,y}) = \frac{\sum_{p,q \epsilon \mathcal{N}_{x,y} \cap \Omega} C(V_{p,q})}{|\mathcal{N}_{x,y}|},$$

where $|\mathcal{N}_{x,y}|$ is the total number of augmented voxels in $\mathcal{N}_{x,y}$. At the beginning, the confidence of each augmented voxel is assigned 1 if its intensity and range values are filled and 0 if the range value is unknown. This confidence term $C(V_{x,y})$ may be thought of as a measurement of the amount of reliable information surrounding the voxel $V_{x,y}$. Thus, as we reconstruct, we synthesize first those voxels whose neighborhood has more of their voxels already filled, with additional preference given to voxels that were synthesized early on. The data term $D(V_{x,y})$ is computed using the

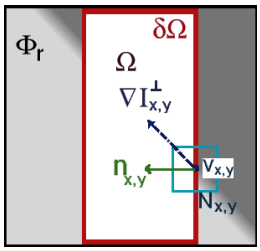

Figure 2: The diagram shows how priority values are computed for each voxel $V_{x,y}$ on $\delta\Omega$. Given the neighboorhood of $V_{x,y}$, $n_{x,y}$ is the normal to the contour $\delta\Omega$ of the target region Ω and $\nabla^{\perp}_{I_{x,y}}$ is the isophote (direction and range) at voxel location x, y.

available range data in the neighborhood $\mathcal{N}_{x,y}$, as follows (see Fig. 2):

$$D(V_{x,y}) = \frac{\alpha}{|\nabla^{\perp}_{I_{x,y}} \cdot n_{x,y}|}. where$$

α is a normalization factor (e.g. $\alpha = 255$, in a typical gray-level image), $n_{x,y}$ is a unit vector orthogonal to the boundary $\delta\Omega$ at voxel $V_{x,y}$. This term reduces the priority of a voxel in whose neighborhood an isophote "flows" into, thus altering the sequencing of the extrapolation process. This term plays an important role in our algorithm because it prevents the synthesis of voxels lying near a depth discontinuity. Note, however, that it does not explicitly alter the probability distribution associated the voxel (except by deferring its evaluation), and thus has only limited risk for the theoretical correctness of the algorithm.

Once all priority values of each augmented voxel on $\delta\Omega$ have been computed, we find the voxel with the highest priority. We then use our MRF model to synthesize its depth value. After a voxel has been augmented (i.e. it has intensity and range data), the confidence of the $C(V_{x,y}) = C(V_{k,l})$, i.e. it is assigned the confidence of the voxel which most resemble the neighborhood of $V_{x,y}$ (see Eq. 2).

4 Experimental Results

In this section we show experimental results conducted on data acquired in a real-world environment. We use ground truth data from a widely available database [1] [18] which provides color images with complex geometry and pixel-accurate ground-truth disparity data. We also show preliminary results on data collected by our mobile robot, which has a video camera and a laser range finder mounted on it. 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 allows us to compare the quality of our reconstruction with what is actually in the scene.

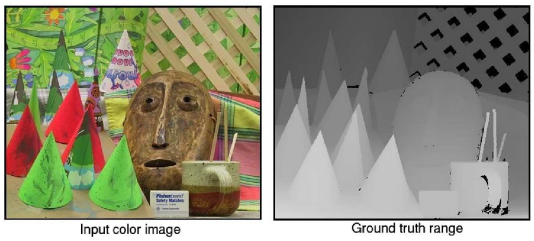

Figure 3: The input intensity image and the associated ground truth range. Since the unknown data are withheld from genuine ground truth data, we can estimate our performance.

Our first set of experiments is interesting because it resembles what is obtained by sweeping a one-dimensional LIDAR sensor. We show different subsamplings on the same range image in order to compare the results. Figure 3 displays the input color intensity image and the ground truth range image from where we hold back the data to simulate the samples. In Figure 4, two experiments are shown. The first column displays the initial range data. The percentage of unknown range from top to bottom are 65% and 62%, respectively. The last column show the synthesized range images when incorporating isophote constraints to our algorithm. For comparison purposes, the middle column shows the synthesized range images without using information about the isophotes. The regions enclosed by the red rectangles show where our algorithm performed poorly without using isophote constraints. The mean absolute residual errors (MAR) in the grey-level range (i.e. 0 for no error and 255 for maximum error) are, from top to bottom, 10.5 and

[1] http://www.middlebury.edu/stereo

12.2, when no using isophote constraints compared to 6.5 and 7.3, when using isophote constraints.

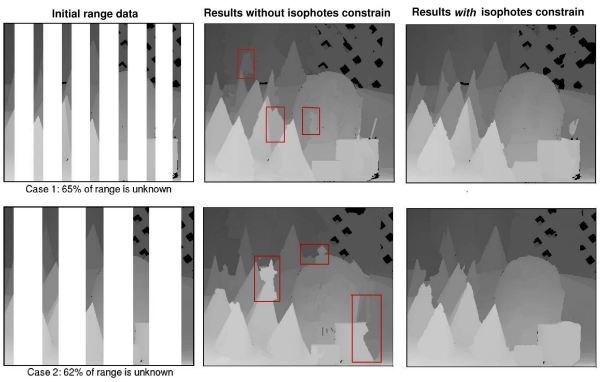

Figure 4: Results on real data. Two cases are shown. The initial range images are in the first column with their percentage of unknown range indicated below each. To compare results, the middle column shows the synthesized range images without using isophote information and the last column show the improved synthesized results with the incorporation of isophote constraints.

It can be seen that our algorithm was able to capture the underlying structure of the scene by being able to reconstruct object boundaries efficiently, even with the small amount of range data given as an input.

We now show another set of experiments in Figure 5. The first row show the input color intensity image and its associated ground truth range (for comparison purposes). Three cases of subsampling are shown in the subsequent rows. The percentage of unknown range are 79%, 70% and 62%, respectively. The MAR errors are 5.94, 5.44 and 7.54, respectively.

Once again, our algorithm was capable of reconstructing the whole range map. However, it can be seen that the reconstruction was not good in regions containing surfaces sloping away (for example walls). This is due to the fact that the very limited amount of input range does not cover much of the changes in depth, and our algorithm fails by assigning already identical depth values instead of different depths at each point. Therefore, 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. As it is now, our method requires that the input range measurements in the form of clusters of measurements scattered over the image. This form of sampling is best since it allows local statistics to be computed, but also provides boundary conditions at various locations in the image. However, clumps per se at not available from most laser range scanners. To solve this problem, we have incorporated surface normal information from voxels with known range values to generate new depth

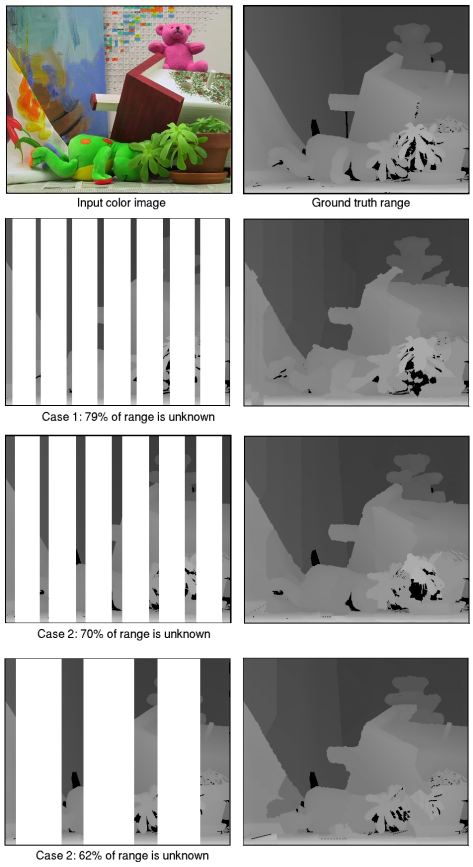

Figure 5: Results on real data. In the first row are the input color image and ground truth range. The subsequent rows show three cases, the initial range images are in the left column and the synthesized results after running our algorithm, in the right column.

values accordingly. In the next section we give details about the computation of surface normals from the available range and how we incorporate this information to our method.

5 Computing surface normals

Surface normal inference is accomplished using a standard plane fitting approach based on an assumption of local bilinearity. The first step is to fit a plane to the points r_i which lie on the $m \times m$ neighborhood of every range point $R_{k,l} \epsilon \Phi_r$. The normal vector to the computed plane is the eigenvector associated with the smallest eigenvalue of the $m \times m$ matrix $A = \sum_i^N ((r_i - c)^T \cdot (r_i - c))$, where c is the *center of gravity* or *centroid* of the set of range points r_i. The smallest eigenvalue ev_s of the matrix A is a measure of the quality of the fit, expressing the deviation of the range points r_i from the fitted plane. We assign normals only to those range points $R_{k,l}$ whose deviation is below a given threshold. In our experiments we use a neighborhood

of size 5×5 and a threshold for the plane fitness of 0.1. A confidence value related to the normal computation is assigned to each augmented voxel $V_{k,l}$. This confidence value is calculated based on the smallest eigenvalue: $1/ev_s$. The incorporation of surface normal information to our algorithm is simple. We just modify the similarity measure between neighborhoods, so that it is based now on the normals instead on the range values. This similarity is computed only when comparing neighborhoods of augmented voxels having already normal information assigned, otherwise it is done as before. A *new* depth value $R_{x,y}$ of the augmented voxel $V_{x,y}$ is computed using the following equation:

$$R_{x,y} = \frac{\mathbf{n} \cdot P - \mathbf{n}_x x - \mathbf{n}_y y}{\mathbf{n}_z},$$

where P are the (x, y, z) coordinates of the augmented voxel $V_{k,l}$ whose neighborhood most resemble the neighborhood of $V_{x,y}$ and $\mathbf{n}$ is the associated normal vector at voxel $V_{k,l}$.

We run some experiments incorporating the surface normal information. In order to compare the results, Figure 6 shows the synthesized range image for Case 3 of Figure 5.

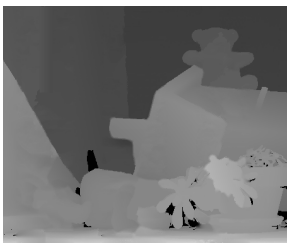

Figure 6: Results when using surface normal information. The initial range data is the Case 3 of Figure 5.

Another experiment is shown in Figure 7. The left image is the initial range data and the input intensity image is that of Figure 3. In order to compare the results, the middle image shows the synthesized range image using the algorithm of the previous section and the right image displays the synthesized range image when incorporating surface normal information to our algorithm.

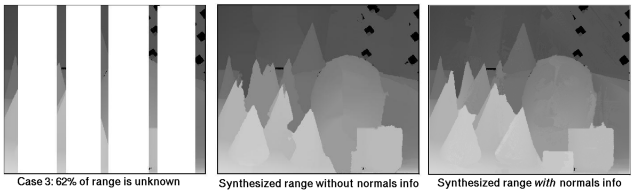

Figure 7: Comparison results when using surface normal information from the initial range data.

These preliminary results show that our method can accomplish the propagation of geometric structure when normal information, from the neighborhoods to be compared with, are available. However, there are some regions where this propagation was not achieved. A key factor in the computation of the normals is the size of the neighborhood containing the range points to fit the plane, which in turn depends on the amount of initial range and on the types of surfaces captured by this initial range.

We now show some preliminary results on data collected in our own building. We use a mobile robot with a video camera and a laser range finder mounted on it, to navigate the environment. For our application, the laser range finder was set to scan a 180 degrees field of view horizontally and 90 degrees vertically. As it was mentioned previously, the input intensity and available range data needs to be already registered. Range and intensity are different type of data, their sampling resolution are not the same. We achieved the registration of the intensity and range data in a semi-automatic way, by using crosscorrelation on the video frames and then manually selecting those corresponding regions from the range image. Details about this registration step is not in the scope of this paper. We are currently seeking to have a fully automatic way of registering both type of data.

Figure 8 shows the input intensity and the ground truth range data (for comparison purposes) on the first row. The second row displays the input range image with 66% of unknown range and the synthesized range data after running our algorithm. It can be seen that the whole depth map of the scene was recovered. Surface normal information was of great use to smoothly generate the new depth values on the walls, floor and ceiling.

6 Summary and Conclusions

In this paper we have presented an approach to depth recovery that allows good quality scene reconstruction to be achieved in real environments using only monocular image data and a limited amount of range data. This methodology is related to extrapolation and interpolation methods and is based on the use of learned Markov models.

We have improved our previous results by adding information about linear structures (isophotes) from the available range in order to drive the reconstruction on the boundary of objects. In addition, we have incorporated surface normal information in the reconstruction process. The preliminary results demonstrate that the fidelity of the reconstruction is improved. Perhaps more important, the robustness of the reconstruction algorithm regarding incomplete input data should be substantially improved. On the other hand, the computation of the surface normals is not an easy task by itself since it depends on the amount of initial range data. There are also some parameters that need to be carefully selected, such as the size of the neighborhood to fit a plane.

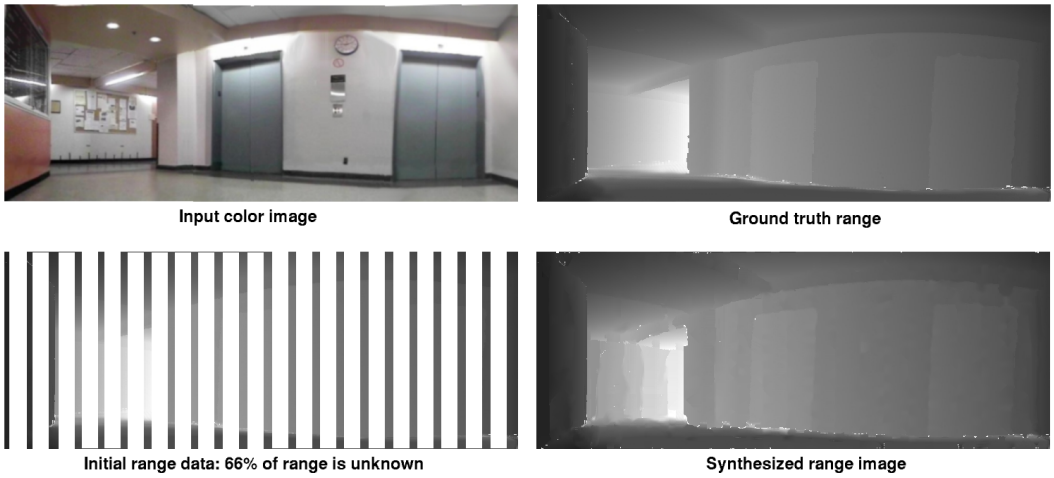

Figure 8: Results on real data collected from our mobile robot.

However, the results shown here demonstrate that is a viable option to obtain a good model of the environment.

References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

[2] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. Simultaneous structure and texture image inpainting. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2003.

[3] A. Criminisi, P. Perez, and K. Toyama. Object removal by examplar-based inpainting. In *IEEE CVPR*, 2003.

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

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

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

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

[8] A. 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.

[9] W. Freeman, E. Pasztor, and O. Carmichael. Shape recipes: scene representations that refer to the image. *Vision Sciences Society Annual Meeting*, pages 25–47, 2003.

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

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

[12] B. Horn and M. Brooks. *Shape from Shading*. MIT Press, Cambridge Mass., 1989.

[13] A. Lee, K. Pedersen, and D. Mumford. The complex statistics of high-contrast patches in natural images, 2001. private correspondence.

[14] 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.

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

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

[17] 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.

[18] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *IEEE CVPR*, volume 1, pages 195–202, 2003.

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

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

[21] A. Torralba and W. Freeman. Properties and applications of shape recipes. In *IEEE CVPR*, 2003.

[22] L. Torres-Méndez and G. Dudek. Reconstruction of 3d models from intensity image and partial depth. 2004. To appear in the AAAI Proceedings, July 24-29, San Jose, California.

[23] L. Torres-Méndez and G. Dudek. Statistical inference and synthesis in the image domain for mobile robot environment modeling. page 8. IEEE Press, September 2004. To appear in the Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS).

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