

Reconstruction of 3D Models from Intensity Images and Partial Depth

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 probabilistic inference of geometric structures from images. Specifically, of synthesizing range data to enhance the reconstruction of a 3D model of an indoor environment by using video images and (very) partial depth information. In our method, we interpolate the available range data using statistical inferences learned from the concurrently available video images and from those (sparse) regions where both range and intensity information is available. The spatial relationships between the variations in intensity and range can be efficiently captured by the neighborhood system of a Markov Random Field (MRF). In contrast to classical approaches to depth recovery (i.e. stereo, shape from shading), we can afford to make only weak prior assumptions regarding specific surface geometries or surface reflectance functions since we compute the relationship between existing range data and the images we start with. Experimental results show the feasibility of our method.

Introduction

This paper presents an algorithm for estimating depth information from a combination of color (or achromatic) intensity data and a limited amount of known depth data. Surface depth recovery is one of the classical standard vision problems, both because of its scientific and pragmatic value. The problem of inferring the 3D layout of space is a critical problem in robotics and computer vision, and is often cited as a significant cognitive skill. In the vision community, solutions to such “shape from X” problems are often based on strong prior assumptions regarding the physical properties of the objects in the scene (such as matte or Lambertian reflectance properties). In the robotics community, such depth inference is often performed using sophisticated but costly hardware solutions.

While several elegant algorithms for depth recovery have been developed, the use of laser range data in many applications has become commonplace due to their simplicity and reliability (but not their elegance, cost or physical robustness). In robotics, for example, the use of range data combined with visual information, has become a key methodology, but 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. 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.

We seek to reconstruct suitable 3D models from sparse range data sets while simultaneously facilitating the data acquisition process. We present an efficient algorithm to estimate 3D dense data from a combination of video images and a limited amount of observed range data. It has been shown in (Lee, Pedersen, & Mumford 2001) 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 remains prohibitive on most mobile platforms. Please note 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 method is based on learning a statistical model of the (local) relationship between the observed range data and the variations in the intensity image and use this model to compute unknown depth values. This can be regarded as a form of shape-from-shading (depth inference from variations in surface shading) based on statistical learning, although traditional shape-from-shading is quite different from this approach in its technical details. Here, we approximate the *composite* of range and intensity at each point as a Markov process. Unknown depth values are 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 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.

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 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 (Geman & Geman 1984; Efros & Leung 1999; Wei & Levoy 2000; Efros & Freeman 2001; Hertzmann *et al.* 2001), but they have never been exploited in tandem. Digital inpainting (Bertalmio *et al.* 2000; 2003; Criminisi, Perez, & Toyama 2003) is quite similar to our problem, although our domain and approach are quite different. In (Baker & Kanade 2002), a learned representation of pixel variation for perform resolution enhancement of face images. The processes employed to interpolate new high-resolution pixel data is quite similar in spirit to what we describe here, although the application and technical details differ significantly. The work in (Torralba & Freeman 2003; Freeman, Pasztor, & Carmichael 2003) on learning the relationships between intrinsic images is also related.

In prior work (Torres-Méndez & Dudek 2002), we performed reconstruction by inferring depth values using predetermined schedule over space, essentially walking a spiral from the boundary of a region towards the center. We have observed that reconstruction across depth discontinuities is often problematic as there is comparatively little constraint for probabilistic inference at these locations. Further, such locations are often identified with edges in both the range and intensity maps. Based on these observations, we have developed two important algorithmic improvements that have a dramatic effect on the quality of the results. In the present work, we modify the reconstruction sequence to first recover the values of those augmented voxels for which we can make the most reliable inferences. This leads to two critical algorithmic refinements: (1) as we recover augmented pixels we defer the reconstruction of augmented voxels close to intensity or depth discontinuities as much as possible, and (2) as we reconstruct we select those voxels for reconstruction that have the largest degree of boundary constraint (aside from those deferred by condition (1)).

Most of prior work on synthesis of 3D environment models uses one of either photometric data or geometric data (Debevec, Taylor, & Malik 1996; Hilton 1996; Fitzgibbon & Zisserman 1998) to reconstruct a 3D model of an scene. For example, in (Fitzgibbon & Zisserman 1998), a method is proposed to 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.

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 methods (Horn & Brooks 1989; Oliensis 1991) 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 (Pulli *et al.* 1997; El-Hakim 1998; Sequeira *et al.* 1999; Levoy *et al.* 2000; Stamos & Allen 2000) have obtained promising results. The work in (Pulli *et al.* 1997) addresses 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 (Sequeira *et al.* 1999) 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 have inherent geometric constraints, such as planar surfaces.

Methodology

Our objective is to infer a dense range map from an intensity image and a very sparse initial range data. At the outset, we assume that the intensity and range data 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.

We observed in previous experiments that the order of reconstruction highly influences the final result. In those experiments, we used a spiral-scan ordering, whose main disadvantage was the strong dependence from the previous assigned depth value. In the present work, our reconstruction sequence is to first recover the values of those locations for which we can make the most reliable inferences, so that as we reconstruct, we select those voxels for reconstruction that have the largest degree of boundary constraint. We have also observed that reconstruction across depth discontinuities is often problematic as there is comparatively little constraint for probabilistic inference at these locations. Further, such locations are often identified with edges in both the range and intensity maps. In this work we have incorporated edge information and, as we recover augmented voxels, we defer the reconstruction of augmented voxels close to intensity or depth discontinuities as much as possible.

The Algorithm

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). Let Ω be the area of unknown range (i.e. the region to be filled). In our algorithm, the depth value r of each augmented voxel $V_i \in \Omega$ is synthesized one at a time. We based our reconstruction sequence on the amount of reliable information surrounding the augmented voxel whose depth value is to be estimated, and also on the edge information. Thus, for each augmented voxel V_i , we count the number of neighbor voxels (in their 8-neighborhood system) with already assigned range and intensity. We start by synthesizing those augmented voxels which have more of their neighbor voxels already filled, leaving to the end those with an edge passing through them. After a depth value is estimated, we update each of its neighbors by adding 1 to their own neighbor counters. We then proceed to the next group of augmented voxels to synthesize until no more augmented voxels in Ω exist.

MRFs for Range Synthesis

Markov Random Fields (MRFs) are used here as a model to synthesize range. Formally, an augmented voxel is defined as $\mathbf{V} = (\mathbf{I}, \mathbf{E}, \mathbf{R})$, where \mathbf{I} is the matrix of known pixel intensities, \mathbf{E} is a binary matrix (1 if an edge exists and 0 otherwise) and \mathbf{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 \mathbf{I} , thus giving us a registered range image \mathbf{R} and intensity image \mathbf{I} . Let $Z_m = (x, y) : 1 \leq x, y \leq m$ denote the m integer lattice (over which the images are described); then $\mathbf{I} = \{I_{x,y}, (x, y) \in Z_m\}$, denotes the gray levels of the input image, and $\mathbf{R} = \{R_{x,y}, (x, y) \in Z_m\}$ denotes the depth values. We model \mathbf{V} as an MRF. Thus, we regard \mathbf{I} and \mathbf{R} as a random variables. For example, $\{\mathbf{R} = r\}$ 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) $(x, y) \in \mathcal{N}_{k,l} \iff (k, l) \in \mathcal{N}_{x,y}$. 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),

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

The choice of \mathcal{N} together with the conditional probability distribution of $P(\mathbf{I} = i)$ and $P(\mathbf{R} = r)$, 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.,

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

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 $\|\cdot\|$ 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.

Experimental Results

In this section we show experimental results conducted on data acquired in a real-world environment. We use ground truth data from two widely available databases. The first database¹ provides real intensity (reflectance) and range images of indoor scenes acquired by an Odetics laser range finder mounted on a mobile platform. The second database² 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.

¹<http://marathon.csee.usf.edu/range/Database.html>

²<http://cat.middlebury.edu/stereo/newdata.html>

This allows us to compare the quality of our reconstruction with what is actually in the scene. In the following, we will consider several strategies for subsampling the range data. In our experiments, we use the Canny edge detector (Canny 1986) for extracting the edges.

Arbitrary shape of unknown range data

The first type of experiment involves the range synthesis when the unknown range data is of arbitrary shape. In particular, we show how the shape that contains the unknown range influences their estimation. In Figure 1a, two input range images (the two left images), and the input intensity with its corresponding edge information (the two right images), are given. The percentage of the missing area (shown in white) of both range images is 41.5% (6800 pixels). The perimeters however are different. Fig. 1b shows the synthesized range images (the two left images) and the ground truth range image (the right image) for comparison purposes.

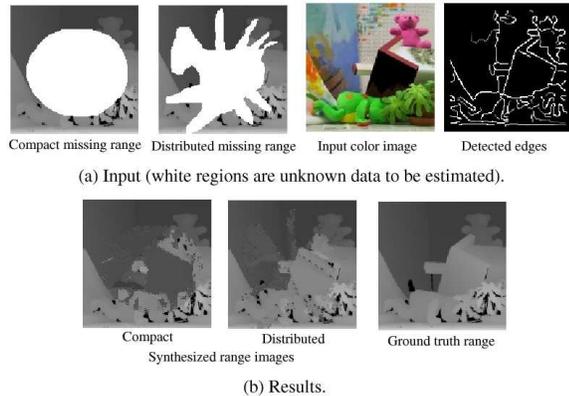


Figure 1: Results on two different shapes of unknown range with same area: 6800 pixels (41.5% of total image).

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 intensity give more information about the geometry of the scene. In other words, the sample spans a broader distribution of range-intensity combinations.

Range measurements with variable width along the x - and y - axis

This type of experiment involves the range synthesis when the initial range data is a set of stripes with variable width along the x - and y -axis of the intensity image. In the following experiments, 61% of the total area is unknown range. We conducted experiments on 32 images of common scenes found in a general indoor man-made environment using this case of subsampling. The smoothing parameter for edge detection was set to 0.8 in all examples. Due to space limitations, we are only showing 4 examples in Figure 2. The absolute value of each error is taken and the mean of those values is computed to arrive at the mean absolute residual (MAR)

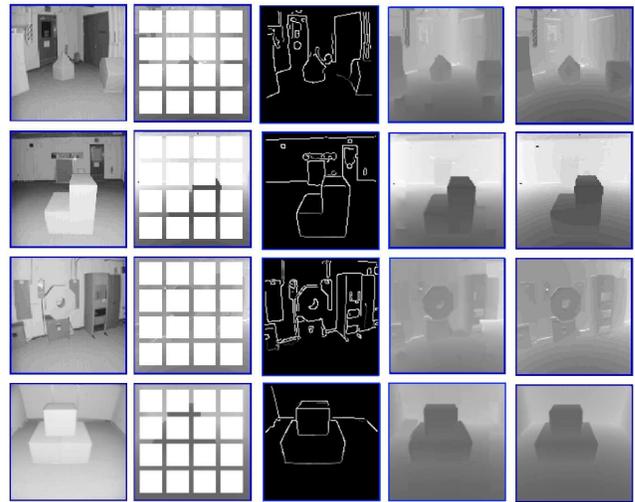


Figure 2: Examples on real data. The first two columns are the input intensity and range data, respectively. The area of missing range is 61%. The third column shows the edges detected in the input intensity. The synthesized results are shown in the fourth column and the ground truth range images are displayed in the last column for visual comparison.

error. The MAR errors from top to bottom are shown in Table 1. The approximated size of each scene is also given. We normalize the MAR error by dividing it by the scene size. This normalized MAR measure is a better indication of how large the error is according to the scene size. Computation time for these results using non-optimized code, is on the order of minutes on a generic PC's.

	Ex. 1	Ex. 2	Ex. 3	Ex. 4
MAR Error (cms)	8.58	13.48	11.39	7.12
Approx. scene size (cms)	600	800	500	400
MAR/Depth	0.017	0.021	0.024	0.048

Table 1: MAR errors of the cases shown in Figure 2.

We now show how color information can improve the synthesized results. Figure 3 displays in the first row, the grayscale and color images of the same scene, and to their right the input range data. The percentage of missing range is 61%. The size of the neighborhood is set to be 5×5 pixels. The synthesized results is shown in the second row together with the ground truth data for comparison purposes. It can be seen that there are some regions where color information may help in the synthesis process. For example, the chimney in the center of the image is separated from the background since they have different colors. This is hardly noticeable in the grayscale image.

Some preliminary results on data collected in our own building is presented next. 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.

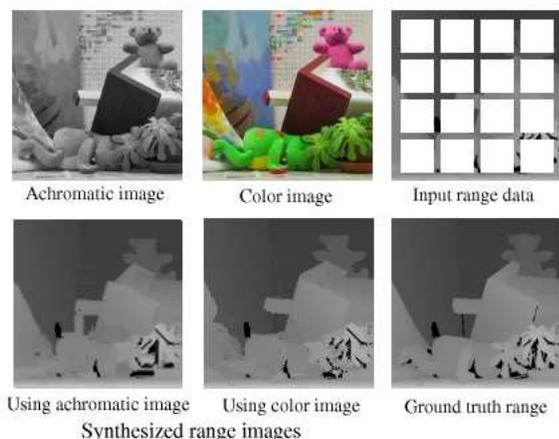


Figure 3: Results on achromatic and color images.

Figure 4 shows experimental results for a case where the input range data is a set of stripes along the x- and y-axis. The input intensity and the ground truth range data (for comparison purposes) are shown on the first row. Two examples are displayed in the second and third row. The left images show the input range images and the right images the synthesized range data after running our algorithm. For the first example, the area of unknown range is 62% of the total image and, for the second example, is 60%.

In general, the synthesized results are good in both cases. Our algorithm was capable of recovering the whole range of the image. There are, however, regions where the synthesis was not good. For example, regions containing surfaces that slope and whose surrounding neighborhoods lack of substantial intensity-range combinations to capture the underlying surface. The reconstruction in those regions becomes a difficult task since it is done from already seen depth values. A solution to this problem, which we are currently working on, is to use surface normal information to generate new depth values according to the type of surface being reconstructed. However, the results presented here demonstrate that this is a viable option to facilitate environment modeling.

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. The method requires both an intensity images and set of partial range measurements input. In fact, the input range measurements are most effective if they are provided 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. While clumps per set are not available from most laser range scanners, swaths of data can be readily and efficiently extracted using laser scanners.

When we use color images in the reconstruction process, it appears that the fidelity of the reconstruction is somewhat

improved over achromatic images. This appears to be due to the fact that the color data provides tighter constraint over where and how the interpolation process should be applied. At the same time, the higher dimensionality of the Markov Random Field model for color images may make the reconstruction problem more difficult in some cases.

We have also evaluated the performance of the reconstruction scheme using intensity data from a simple CCD camera and range data from a laser time of flight scanner (LIDAR) mounted near the camera. While these two sensors are mount close together, there remains an unavoidable stereo disparity between the two sensors. This is probably responsible for portion of the reconstruction error we observed. While this is slightly problematic it seems amenable to various solution techniques; in fact, it may be possible to exploit this disparity data as an additional form of stereo.

Critical to the performance of this method is the statistical similarity of the regions being reconstructed and the portions of the image used to define the Markov model. An open question is how to validate this statistical similarity which could be useful both to control and to validate the reconstruction process.

Acknowledgements

We would like to thank the CESAR lab at Oak Ridge National Laboratory in Tennessee and the Stereo Vision Research Group at Middlebury College for making their range image databases available through their websites.

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) and NSERC for ongoing funding.

References

- Baker, S., and Kanade, T. 2002. Limits on super-resolution and how to break them. *IEEE Trans. on PAMI* 24(9):1167–1183.
- Bertalmio, M.; Sapiro, G.; Caselles, V.; and Ballester, C. 2000. Image inpainting. In *SIGGRAPH*, 417–424.
- Bertalmio, M.; Vese, L.; Sapiro, G.; and Osher, S. 2003. Simultaneous structure and texture image inpainting. In *IEEE CVPR*.
- Canny, J. 1986. A computational approach to edge detection. *IEEE Trans. on PAMI* 8(6):679–698.
- Criminisi, A.; Perez, P.; and Toyama, K. 2003. Object removal by exemplar-based inpainting. In *IEEE CVPR*.
- Debevec, P.; Taylor, C.; and Malik, J. 1996. Modeling and rendering architecture from photographs: A hybrid geometry and image-based approach. In *SIGGRAPH*, 11–20.
- Efros, A., and Freeman, W. 2001. Image quilting for texture synthesis and transfer. In *SIGGRAPH*, 1033–1038.
- Efros, A., and Leung, T. 1999. Texture synthesis by non-parametric sampling. In *ICCV (2)*, 1033–1038.
- El-Hakim, S. 1998. A multi-sensor approach to creating accurate virtual environments. *Journal of Photogrammetry and Remote Sensing* 53(6):379–391.

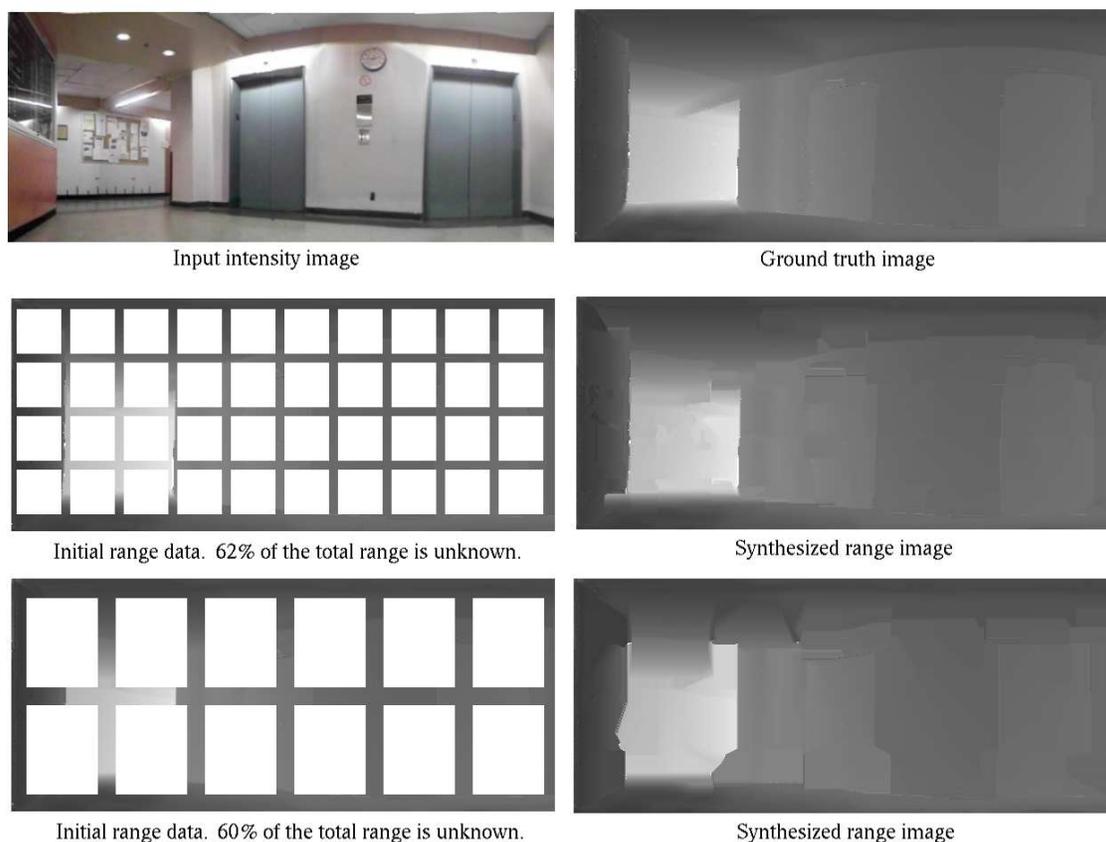


Figure 4: Results on real data collected from our mobile robot.

Fitzgibbon, A., and Zisserman, A. 1998. Automatic 3d model acquisition and generation of new images from video sequences. In *Proceedings of European Signal Processing Conference*, 1261–1269.

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

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

Hertzmann, A.; Jacobs, C.; Oliver, N.; Curless, B.; and Salesin, D. 2001. Images analogies. In *SIGGRAPH*.

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

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

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

Levoy, M.; Pulli, K.; Curless, B.; Rusinkiewicz, S.; Koller, D.; Pereira, L.; Ginzton, M.; Anderson, S.; Davis, J.; Ginsberg, J.; Shade, J.; and Fulk, D. 2000. The digital michelangelo project: 3d scanning of large statues. In *SIGGRAPH*.

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

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

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

Stamos, I., and Allen, P. 2000. 3d model construction using range and image data. In *IEEE CVPR*.

Torralla, A., and Freeman, W. 2003. Properties and applications of shape recipes. In *IEEE CVPR*.

Torres-Méndez, L., and Dudek, G. 2002. Range synthesis for 3d environment modeling. In *IEEE Workshop on Applications of Computer Vision*, 231–236.

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