# Inter-Image Statistics for Scene Reconstruction

Luz A. Torres-Méndez, Gregory Dudek and Paul Di Marco
Center for Intelligent Machines, McGill University
Montreal, QC, Canada
{latorres,dudek,dimarco}@cim.mcgill.ca

## Abstract

*This paper developed prior work which incrementally completes a sparse depth map based on inter-image statistics information. In that prior work, we have observed that pixel ordering of the incremental recovery is critical to the quality of the final results. In this paper we demonstrate improved performance using an information-driven recovery policy to determine this ordering. We have also observed that the reconstruction across depth discontinuities was often problematic as there was comparatively little constraint for probabilistic inference at those locations. Further, such locations are often identified with edges in both the range and intensity maps. We address this problem by deferring the reconstruction of voxels close to intensity or depth discontinuities, leading to improved results. We also show that color information can improve reconstruction quality. Experimental results are presented to demonstrate the quality of the recover and to illustrate some new application domains such as deblurring and underwater scattering compensation.*

## 1. Introduction

In this paper we consider the use of statistical models to transfer information between different scene representations. In particular, we consider a collection of intrinsic images of a scene [2] in terms of their joint statistics (by intrinsic images, we mean retinotopic maps of scene properties, such as irradiance and depth). We make the assumption that this join space obeys the Markov property and can be modeled using a Markov Random Field. In particular, in this paper we consider the joint space composed of image intensity and depth. While the intensity map for a scene does not strictly obey the Markov property, this assumption that it does, seems acceptable and has been used extensively.

Given that we can compute the statistical properties of the Markov Random Field over range and intensity, we can then use it to make various estimates. In particular, in the absence of complete observations, we can fill in the missing data by making probabilistic guesses based on the statistics of the joint space and the partial data we do have.

In this paper we use such an approach to infer a dense range map given only sparse initial estimates. This is a problem of substantial practical importance since the acquisition of accurate range data can be slow or 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. Specifically, in many robotics applications sparse range data can be obtained by sweeping a laser range scanner over the surfaces of interest while dense range data may take too long to measure. In such an instance, the partial data allows us to compute the required statistics which can then be used to infer the missing data. In this process, the availability of ubiquitous image data constrains the reconstruction process and makes it feasible.

The same approach also seems applicable to several related problems. In the latter part of this paper we demonstrate its feasibility to image deblurring to compensation for scattering in underwater images (another domain in which we are actively conducting experiments).

## 2. Background

Our work is an instance of the 3D environment modeling problem. Over the last 30 years, this problem has received considerable attention in the computer vision and computer graphics communities and more recently in robotics. In the context of this paper we will consider only a few representative solutions.

It is our belief that at least two types of data are essential to facilitate the reconstruction of a 3D model of an object or scene. One is photometric data that can provide high accuracy on features and edges. The other is range data that provides the geometric information. Intensity images alone, cannot provide complete or accurate 3D measurements on unmarked continuous surfaces, therefore both types of data should be integrated.

In the literature, however, much of the previous work create 3D models directly from photometric data. Some of these methods are based on projective calibration and reconstruction techniques [7, 15, 11, 19]. For example, Fitzgibbon and Zisserman [11] 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 trifocal 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. [19]; 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 it is based on a rather different set of assumptions and methodologies. Such method [16, 18] reconstructs 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 with promising results [20, 10, 21, 17, 22]. In their approach, Pulli et al. [20] measure both color and geometry of real objects, and display 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. 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.

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 [13, 9, 26, 8, 14], but they have never been exploited in tandem.

Digital inpainting [3, 4, 6] is quite 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 quite similar in spirit to what we describe here, although the application and technical details differ significantly. The work by Freeman [23, 12] on learning the relationships between intrinsic images is also related.

Our method bases its reconstruction process on having a small amount of range data and synthetically estimating the areas of missing range by using the current available data. Except for our earlier work [24, 25], we have not found published work dealing specifically with the approach we are taking. In particular, such a method is feasible in man-made environments, which, in general, have inherent geometric constraints, such as planar surfaces.

## 3. Our Statistical Approach for Inferring Depth

We based our approach on earlier work described by Torres-Mendez and Dudek [24]. In that work, the pixel ordering of reconstruction (the order in which we choose the next depth value to synthesize) was determined using a predetermined schedule over space, essentially walking a spiral from the perimeter of a region towards the center. One of the problem with the spiral-scan ordering was the strong dependence on the previously estimated pixel. In the present work, we use an information-driven approach, in which the order of reconstruction is to first recover the depth values of those locations for which we can make the most reliable inferences, so that as we reconstruct we select those pixels for reconstruction that have the largest degree of boundary constraint. We also have observed that the 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. This observation leads to another modification in our reconstruction sequence: as we recover depth values, we defer the reconstruction of those pixels close to intensity or depth discontinuities as much as possible.

The images used in the reconstruction process can be achromatic (black and white) or color. In this paper we compare the reconstructions using these two types of input. It appears that color information improves the reconstruction accuracy. This may 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.

## 4. Algorithm description

Our objective is to compute depth values where only intensity is known. We will do this by incrementally computing a single depth value at a time by using neighboring locations where both range and intensity is available. At the outset, we assume that resolution of the intensity and range data is the same and that they are already registered.

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 regarding two kinds of scene phenomenon: (1) knowledge of when the surface is smooth, and (2) knowledge of when there is a high probability of a variation in depth. In reality, the statistical information implicit in the data may be much more subtle than simply these two types of event, but they illustrate the concept. Our approach learns the required relationships from the observed data, without having to fabricate or hypothesize constraints that might be inapplicable to a particular environment.

### 4.1. The Modified MRF Model

Markov Random Fields (MRFs) are used here as a model to synthesize range. We focus on our development of a set of **augmented voxels 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) $(x, y) \not\in \mathcal{N}_{x,y}$, 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),

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

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.,

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

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 $\| \,.\, \|$ 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.

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. We use the Canny edge detector [5] for extracting the edges from the intensity images. Let $V_p$ be an augmented voxel with unknown range and $\mathcal{N}_p$ be a $3 \times 3$ square window centered at $V_p$ (i.e. we are considering just the 8-closest neighbors). Then, for each augmented voxel $V_p$, we count the number of neighbor voxels with already assigned range and intensity. We start by synthesizing those augmented voxels with the maximum number of

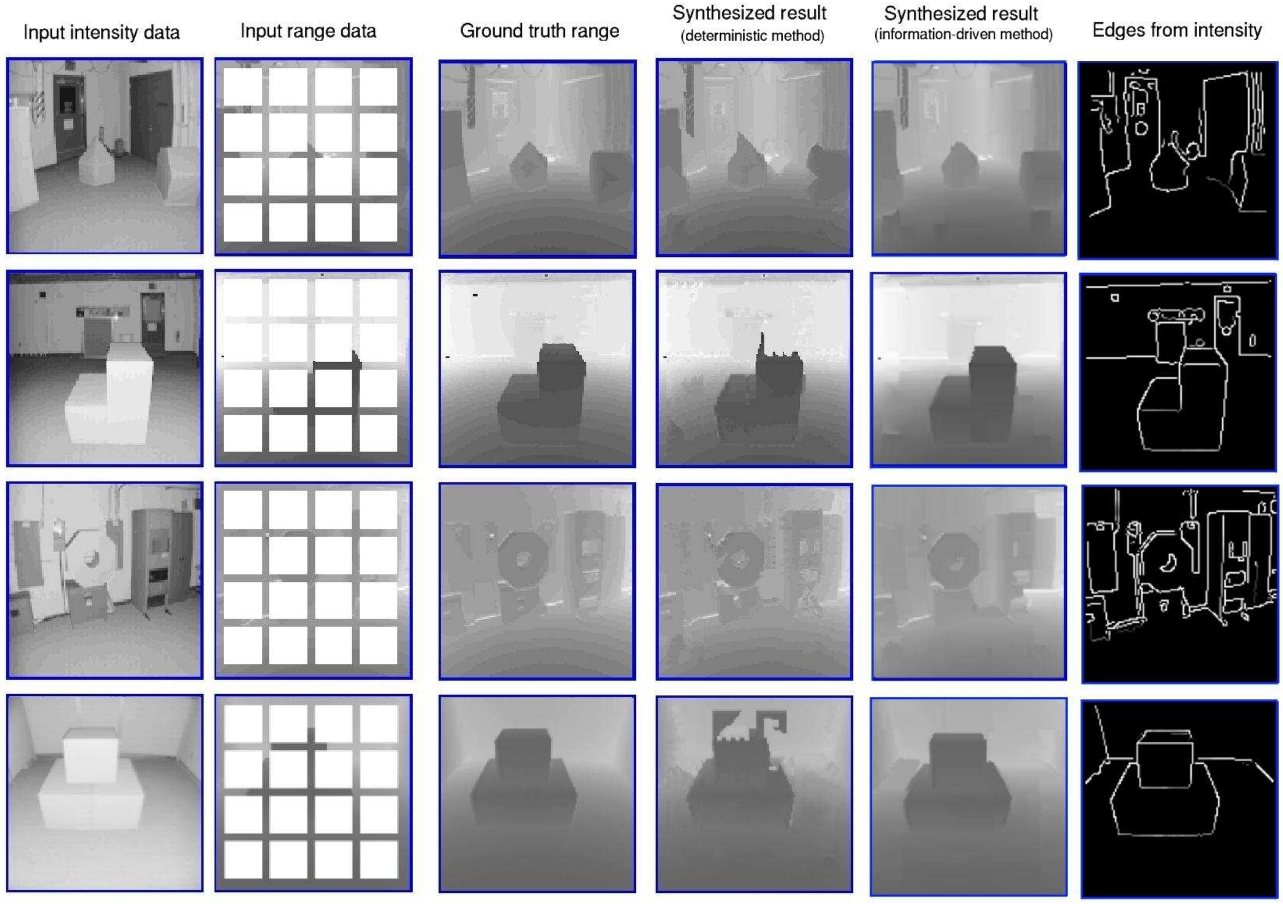


**Figure 1. Comparison of the reconstruction performance using the deterministic and information-driven methods. The first two columns display the input intensity and the input range data where 61% of the total is unknown (the white squares), respectively. The third column shows the ground truth range for comparison results. In the fourth column, the synthesized results using the deterministic method are shown, and the last two columns show the synthesized results using the information driven method and the detected edges from the intensity input images, respectively.**

filled neighbors, 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 exist.

## 5. Experimental Results

We run our improved algorithm on data acquired in a real-world environment. As we did in our earlier work, we use ground truth data from two widely available databases. The first database [1] provides real intensity (reflectance) and range images of indoor scenes acquired by an Odetics laser range finder mounted on a mobile platform. The second database [2] provides color images with complex geometry and pixel-accurate ground-truth disparity data. 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.

In Figure 1, we compare the synthesized results over 4 images using the deterministic (spiral-scan ordering) versus the information-driven approach. The three left columns are, from left to right, the input intensity image, the input range data and the ground truth range (for comparison purposes). The percentage of the unknown range (shown in

[1]http://marathon.csee.usf.edu/range/Database.html

[2]http://cat.middlebury.edu/stereo/newdata.html

white) of all input range images is 61%. The fourth column shows the synthesized results using the deterministic version of our algorithm. Note how the algorithm performed poorly near object edges, where high discontinuities exist, specially in the last image, the synthesis started to be wrong and continue in a spiral fashion because of the order of the reconstruction used. The last two columns show the synthesized results after running our information-driven version of the algorithm and the detected edges from the input intensity images that were used, respectively. Note how our algorithm was able to recover well the geometry of the objects in the scenes. Quantitative results of these experiments are given in Table 1. The absolute value of each error is taken and the mean of those values is computed to arrive at the mean absolute residual (MAR) error. The approximated scene size of each scene is also given.

| MAR Errors (in cms) | | Approx. scene size (in cms) |
|---|---|---|
| Deterministic | Information-driven | |
| 10.40 | 8.58 | 600 |
| 16.58 | 13.48 | 800 |
| 12.16 | 11.39 | 500 |
| 19.17 | 7.12 | 400 |

**Table 1. The input information and MAR errors of the cases shown in Figure 1.**

We now show how color information can improve the synthesized results. Figure 2 displays in the first row, the input images (achromatic and color) and to their right the input range data. The percentage of missing range is 61%. The size of the neighborhood is set to be $5 \times 5$ pixels.The synthesized results after running our algorithm 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. The MAR errors are 7.71 when using grayscale and 6.39 when using color information.

Another example is displayed in Figure 3. As the first example, the first row shows the input images (achromatic and color) and to their right is the input range. For this case, 71% of the total range map is unknown. The left image of second row displays the synthesized result and to its right is the ground truth range for comparison purposes. The MAR errors are 9.41 when using grayscale and 7.14 when using color information. In this case, color was useful in the reconstruction of the cones. It is important to note that this is a difficult scene, in particular because of the many features present on it and the limited input range that is given.

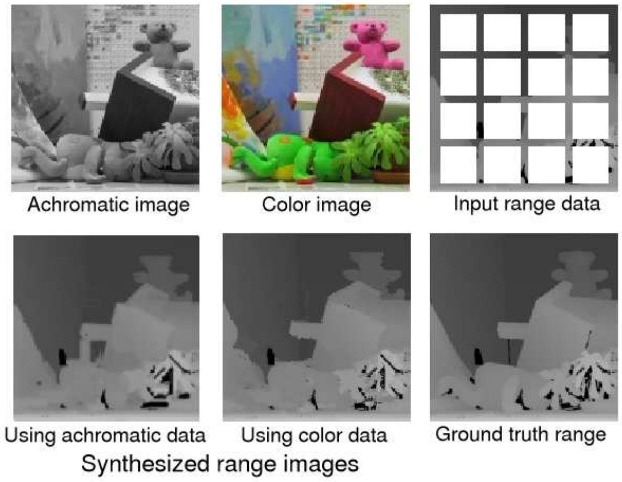


**Figure 2. Results on achromatic and color images. The MAR error is 7.71 when using grayscale and 6.39 when using color information.**

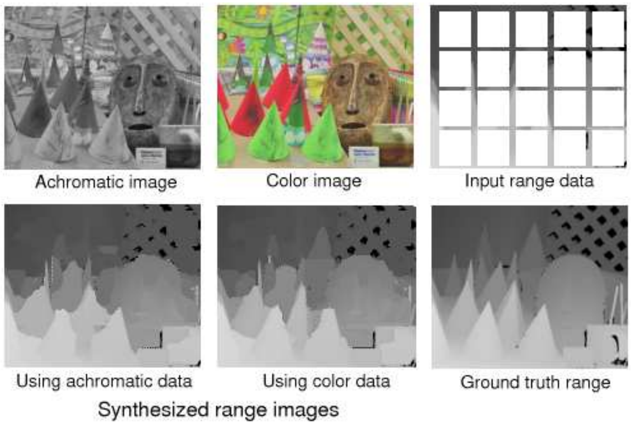


**Figure 3. Results on grayscale and color images. 71% of the total range map is unknown (shown in white). The MAR error when using color is 7.14 compared to 9.41 when using grayscale information.**

### 5.1. Another application: Image Deblurring of Underwater Scenes

Markov Random Fields are commonly used for image restoration. We are currently using our method in the removal of hazing due to optical scattering underwater. In this application, our algorithm learns the statistical relationships between blurred and deblurred pixels in small patches or neighborhoods (usually $3 \times 3$ or $5 \times 5$ pixels) on images

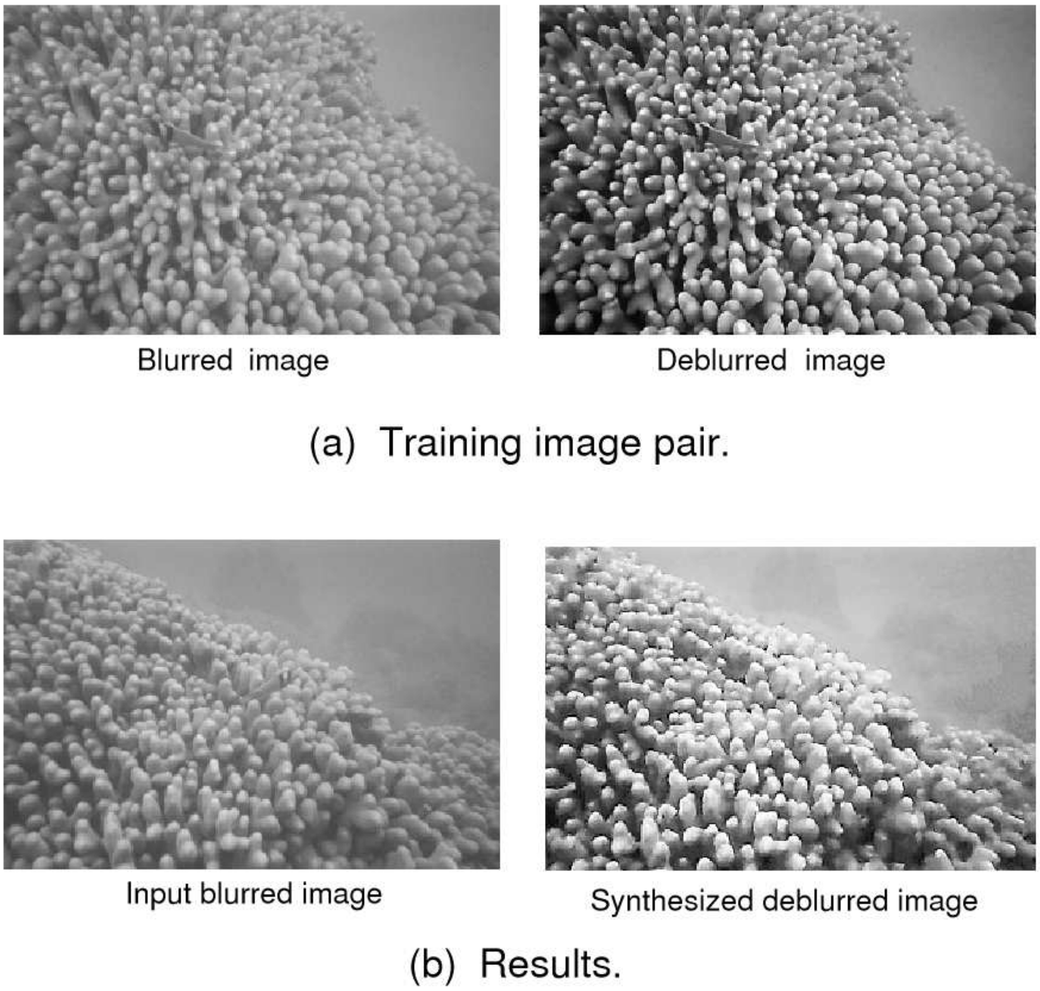


**Figure 4. Results on image deblurring. (a) The blurred and deblurred images to train our algorithm. (b) The left image is the test blurred image and to its right is the resulting deblurred image.**

of the same underwater scene. These images are the training images, then when a new blurred image is input to the algorithm, the underlying statistics already captured in the training sets, are used to estimate what a corresponding deblurred pixel should look like. Given the neighborhood $N_p$ of the pixel $p$ from image $D$ to deblurr, whose neighboring pixels may contain blurred and deblurred pixels, we find the most similar neighborhood $N_q$ of the pixel $q$ from the training images. For the first pixel to deblurr, its neighboring pixels contain only blurred pixels, but as the synthesis progresses, more deblurred pixels are taking into account helping in the quality of the result. The deblurred value of pixel $q$ is assigned to the deblurred value of pixel $p$.

In Figure 4 we show an example. Given the pair of images of a coral reef scene shown in Figure 4a, where the left image is a blurred version of the right image, our algorithm computes their joint statistics for deblurring the left image shown in Figure 4b. The right image of Figure 4b is the resulting deblurred image. We can see that most of the tiny features of the coral reef in the blurred image were efficiently deblurred.

## 5.2. Conclusions

This paper has sumarized and approach to scene reconstruction and range map inference using intensity data. The method is based on learning the statistical relationship between range and intensity using sample data from the image pair to be recovered. It also appears that the method can work well using images pairs from one part of a scene to reconstruct range data from another part of the scene, although those depends critically on the statistical similarity between these regions (i.e. they need to “look similar”).

We have observed that this process can take place using either color or achromatic images, but that slightly better results are obtained for color images. Superficially we can explain this by observing that color images have “more information” in them. More specifically, color images typically contain supplementary subtle cues regarding the distinction between marking and surface boundaries, which are key to the reconstruction process. We have also demonstrated that the use of edge image to alter the reconstruction sequence can substantially improve the quality for the results. Again, this relates to the treatment of surface boundaries, where the reconstruction process is particularly difficult.

Finally, we have observed that the same type of process can also be used for image deblurring. A similar application is the removal of hazing due to optical scattering underwater. This latter application appears especially important for underwater applications. It is also likely that the depth reconstruction process will operate well for underwater images, but our promising preliminary results are difficult to evaluate due to a lack of ground truth images. In the underwater domain, the scattering can assist the inference of depth from intensity, but the complexity of the naturally occurring scene geometry makes the problem challenging.

## 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

[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] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, pages 3–26, 1978.

[3] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proc. ACM Conf. Computer Graphics (SIGGRAPH)*, pages 417–424, July 2000.

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

[5] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[6] A. Criminisi, P. Perez, and K. Toyama. Object removal by examplar-based inpainting. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2003.

[7] 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.

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

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

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

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

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

[23] A. Torralba and W. Freeman. Properties and applications of shape recipes. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2003.

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

[25] L. Torres-Méndez and G. Dudek. Range synthesis for 3d environment modeling. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, page 8, Las Vegas, NV, October 2003. IEEE Press.

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