Range synthesis for 3D Environment Modeling

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

Abstract

In this paper a range synthesis algorithm is proposed as an initial solution to the problem of 3D environment modeling from sparse data. We develop a statistical learning method for inferring and extrapolating range data from as little as one intensity image and from those (sparse) regions where both range and intensity information is available. Our work is related to methods for texture synthesis using Markov Random Field methods. We demonstrate that MRF methods can also be applied to general intensity images with little associated range information and used to estimate range values where needed without making any strong assumptions about the kind of surfaces in the world. Experimental results show the feasibility of our method.

1. Introduction

In this paper, we present an efficient algorithm for extrapolating sparse range data to provide a dense 3D environment model. The extraction of range data from intensity images has been one of the key problems for computer vision, as addressed by numerous "shape-from" methods. Many of these methods provide incomplete data. More prosaically, while laser range sensors (for example based on LIDAR) have become well-established technology they are costly and often return data that is sparse relative to video images (most typically providing only dimensional-stripes of range data). In robotics, for example, the use of range data for navigation and mapping 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.

Using an intensity image and incomplete range data we develop a statistical learning method that infers missing range data from a partial depth map, such as one obtained by sweeping a one-dimensional LIDAR sensor. Our method is novel in that it allows the estimation of the geometry of the scene with only an intensity image and a relatively small amount of range data and without strong *a priori* assumptions on either surface smoothness or surface reflectance. We base our range estimation process on inferring a statistical relationship between intensity variations and existing range data. This is elaborated using a Markov Random Field (MRF) method akin to those used in texture synthesis [6, 2, 17, 3, 7]. In the context of texture synthesis, MRF methods model a texture based on its local and stationary properties. A new texture is generated pixel by pixel in such a way that these two properties are preserved in a small set of spatially neighboring pixels which characterizes every pixel on the texture image.

The organization of this paper is as follows. In the next section, we review some related work. In Section 3, we describe our range synthesis method. Some experimental results are shown in Section 4, and, finally, in Section 5 we give some conclusions and suggestions for future research.

2. Background

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

Over the last decade laser rangefinders have become affordable and available but their application to building full 3D environment models, even from a single viewpoint, remains costly or difficult in practice. In particular, while laser line scanners based on either triangulation and/or time-of-flight are ubiquitous, full volume scanners tend to be much more complicated and 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.

Much of the previous work on environment modeling uses one of either photometric data or geometric data [1, 8, 5, 12] to reconstruct a 3D model of an scene. For example, Fitzgibbon and Zisserman [5] proposed a method that sequentially retrieves the projective calibration of a complete image sequence based on tracking corner and/or line features over two or more images, and reconstructs each feature independently in 3D. Their method solves the feature correspondence problem based on the fundamental matrix and tri-focal tensor, which encode precisely the geometric

constraints available from two or more images of the same scene from different viewpoints. Related work includes that of Pollefeys et. al. [12]; they obtain a 3D model of an scene from image sequences acquired from a freely moving camera. The camera motion and its settings are unknown and there is no prior knowledge about the scene. Their method is based on a combination of the projective reconstruction, self calibration and dense depth estimation techniques. In general, these methods derive the epipolar geometry and the trifocal tensor from point correspondences. However, they assume that it is possible to run an interest operator such as a corner detector to extract from one of the images a sufficiently large number of points that can then be reliably matched in the other images.

Shape-from-shading is related in spirit to what we are doing, but is based on a rather different set of assumptions and methodologies. Such method [9, 11] reconstruct a 3D scene by inferring depth from a 2D image; in general, this task is difficult, requiring strong assumptions regarding surface smoothness and surface reflectance properties. Recent work has considered the use of both intensity data as well as range measurements. Several authors [13, 4, 14, 10, 15] have obtained promising results. Pulli et al. [13] address the problem of surface reconstruction by measuring both color and geometry of real objects and displaying realistic images of objects from arbitrary viewpoints. They use a stereo camera system with active lighting to obtain range and intensity images as visible from one point of view. The integration of the range data into a surface model is done by using a robust hierarchical space carving method. The integration of intensity data with range data has been proposed [14] to help define the boundaries of surfaces extracted from the 3D data, and then a set of heuristics are used to decide what surfaces should be joined. For this application, it becomes necessary to develop algorithms that can hypothesize the existence of surface continuity and intersections among surfaces, and the formation of composite features from the surfaces.

However, one of the main issues in using the above configurations is that the acquisition process is very expensive because dense and complete intensity and range data are needed in order to obtain a good 3D model. As far as we know, there is no method that bases its reconstruction process on having a small amount of intensity and/or range data and synthetically estimating the areas of missing information by using the current available data. In particular, such a method is feasible in man-made environments, which, in general, have inherent geometric constraints, such as planar surfaces.

3. The Algorithm

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

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

3.1. Synthesizing range

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

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

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

model a neighborhood $\mathcal{N}_{x,y}$ as a square mask of size $n \times n$ centered at the voxel location (x, y). This neighborhood is causal, meaning that only those voxels already containing both, intensity and range information are considered for the synthesis process. On the other hand, calculating the conditional probabilities in an explicit form is an infeasible task since we cannot efficiently represent or determine all the possible combinations between augmented voxels with its associated neighborhoods. Therefore, we synthesize a depth value $R_{x,y}$ deterministically by selecting the range value $R_{k,l}$ from the augmented voxel whose data most resemble the (partial) data from location (x, y), i.e.,

$$\underset{(k,l) \in \mathcal{A}}{\textbf{argmin}} \; ||V_{x,y} - V_{k,l}||, \tag{2}$$

where $\mathcal{A}$ is the set of those augmented voxels located at distance d to the augmented voxel to be synthesized. The similarity measure $||.||$ is described over the partial data about locations (x, y) and (k, l) and is calculated as follows,

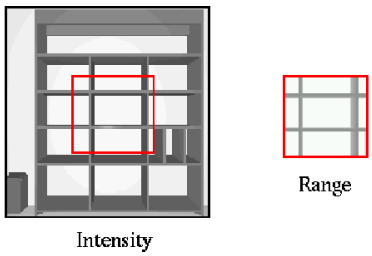
$$\sum_{\vec{v} \in \mathcal{N}} G(\sigma, \vec{v} - \vec{v}_0)[(I_{\vec{v}} - I'_{\vec{v}})^2 + (R_{\vec{v}} - R'_{\vec{v}})^2], \tag{3}$$

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

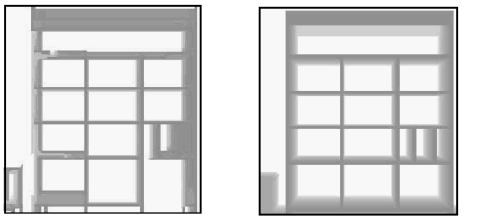
In our algorithm we synthesize one depth value at a time. The order in which we choose the next depth value to synthesize will reflect the final result. In our experiments, depth values are assigned in a spiral-scan ordering, either growing inwards or outwards, depending on the shape of the area to synthesize.

4. Experimental Results

We have tested our algorithm using synthetic and real data. The synthetic data were generated with the 3D rendering package PovRay using natural lighting conditions. In the left side of Figure 1a, a synthetic intensity image of an empty bookshelf is shown; the subwindow in the right side is the associated range image taken from the position indicated by the red rectangle in the intensity image. These images are given as an input to our algorithm. For this case, the size of the neighborhood is set to be 9×9 pixels. The left side of Figure 1b shows the range synthesis results, and to the right, as a way of visual comparison, the complete synthetic range data is shown. It can be seen that the algorithm captures most of the changes involved in the intensity information and are reflected in the range synthesis process.



(a) Input.



(b)

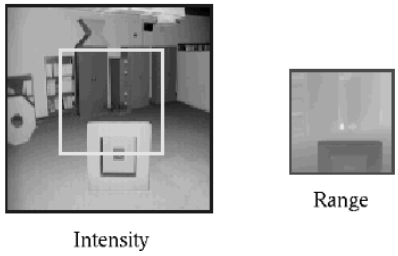
Figure 1. Results of range synthesis on synthetic data. Panel (b) shows the comparison of synthesized range with ground truth for this artificial scene.

Representative results based on data acquired in a real-world environment are shown in Figures 2 through 6. The real intensity (reflectance) and range images of indoor scenes were acquired by an Odetics laser range finder mounted on a mobile platform. Images are 128×128 pixels and encompass a $60^o \times 60^o$ field of view. As with the synthetic 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 allow us to compare the quality of our reconstruction with what is actually in the scene. In the following we will consider two strategies for subsampling the range data.

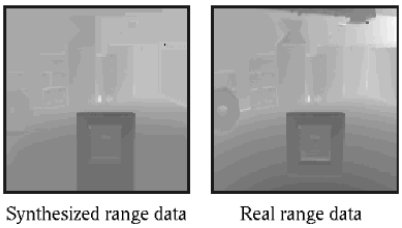
4.1. Limited dense range

The first type of experiment involves the range synthesis when the initial range is a window of size $p \times q$ and at position (r_x, r_y) on the intensity image. Figure 2a shows the intensity image (left) of size 128×128 and the initial range (right), a window of size 64×64, i.e. only the 25% of the total range is known. The size of the neighborhood is 5×5

pixels. The synthesized range data obtained after running our algorithm is shown in the left side of Figure 2b; for purposes of comparison, we show the complete real range data (right side). It can be seen that the synthesized range is very similar to the real range. The Odetics LRF uses perspective projection, so the image coordinate system is spherical. To calculate the residual errors, we first convert the range images to the Cartesian coordinate system (range units) by using the equations in [16]. For this example, the average residual error is 7.98.



(a)

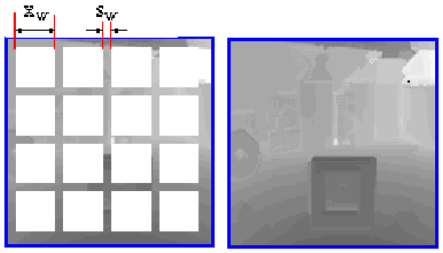


(b)

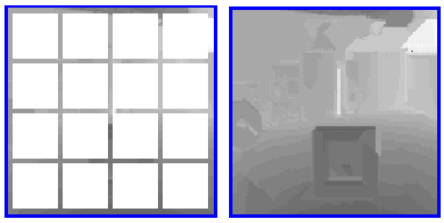
Figure 2. Results on real data. (a) Input. (b) Results comparing synthesized range data to ground truth.

4.2. Sparse range measurements

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



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



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

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

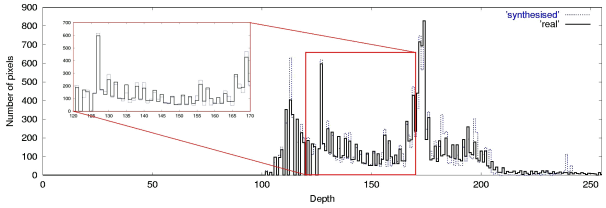


Figure 4. Histogram of pixels at different depth values (scale from 0 to 255) of the original and synthesized range of Figure 3a.

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

Our algorithm was tested on 30 images of common scenes found in a general indoor man-made environment. Two cases of subsampling were used in our experiments. Case 1 is as one of the subsampling previously described, with $r_w = 5$ and $x_w = 25$, applied along $x-$ and $y-$axis.

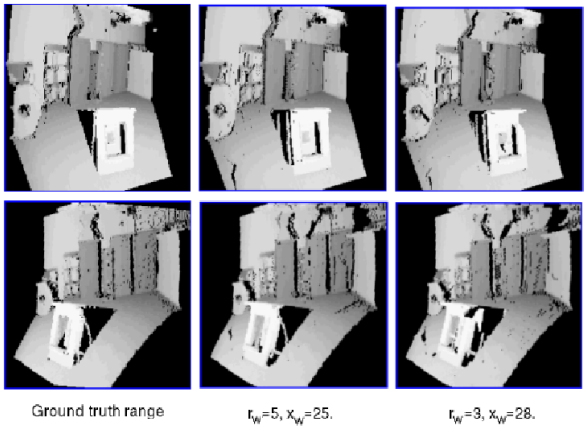


Figure 5. Results in 3D. Two views of the real range (left column) and the synthesized results (middle and right columns) of Figure 3.

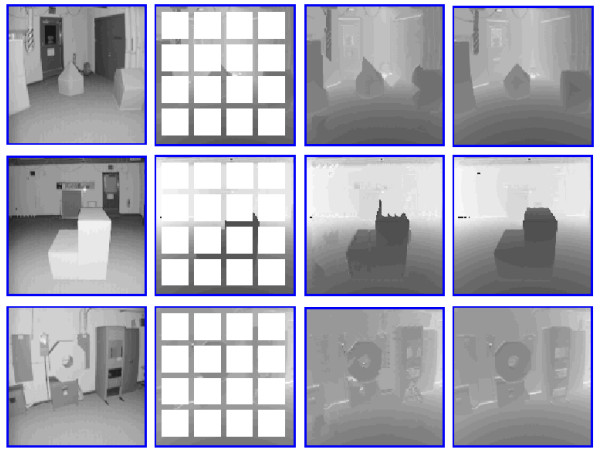
Due to space limitations, we are only showing 3 more examples of this case in Figure 6a, the average residual errors are, from top to bottom, 2.84, 4.53 and 3.32. For Case 2, $r_w = 8$ and $x_w = 22$, but applied only along the $x-$axis. Figure 6b shows 2 examples of this case. Here the average residual errors are 4.17 and 5.25, respectively. Once again, it can be seen that the results are good in both cases. The maximum average residual errors obtained from all 30 test images were for Case I, 6.52 and for Case II, 11.85.

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

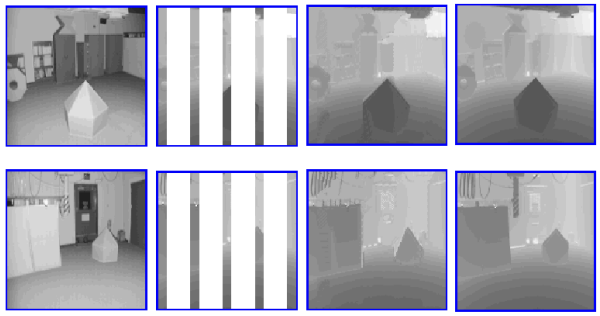
5. Conclusions and Future Work

We have presented an algorithm for recovering 3D geometric data given an intensity image with little associated range information. The approach uses Markov Random Field methods as a model that relates characteristics of the intensity and range data. There are a number of parameters that can greatly influence the quality of the results: the size of the neighborhood used in computing correlations, the amount of initial range and the characteristics captured in that initial range. The characterization of how these parameters effect the results is the subject of ongoing work.

Our approach as described in this paper exploits the statistically observed relationship between the intensities in a neighborhood and range data to interpolate (or extrapolate) the range. While this formalism can explicitly capture local differential geometry, we do not explicitly compute local surface properties, nor does this approach make substantive assumptions regarding surface reflectance functions of surface geometry such as smoothness. The approach does assume that the relationship between intensity and range can be expressed by a stationary distribution; an assumption that could be relaxed. While avoiding strong assumptions about the surfaces in the scene allows greater generality, it also means we do not exploit potentially useful constraint information. In ongoing work, we are examining the incorporation of more elaborate priors and geometric inferences.



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



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

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

Another interesting problem we are exploring on, is that of inferring intensity from range, as opposed to range from intensity. This would permit inference of intensity distributions in cases where surface reflectances were difficult to model (such as on textured of patterned surfaces) and might

serve as an adjunct to more conventional reflectance-based modeling. The approach described here may, in principle, work well; however, difficulties arise because range data does not provide information about what kind of textures are in the intensity image, so additional information should be considered. The following examples shown in Figure 7 illustrate this. The average residual errors, considering the gray levels (0 to 255), are 11.45 and 12.70, respectively.

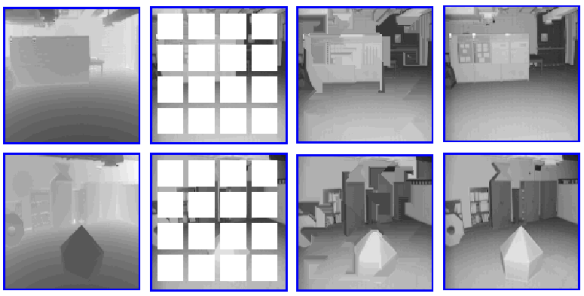


Figure 7. Inferring intensity from range.

Acknowledgements

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

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

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

References

[1] P.E. Debevec, C.J. Taylor and J. Malik, "Modeling and Rendering Architecture from Photographs: A hybrid geometry and image-based approach," *Proceedings of SIGGRAPH'96*, pp. 11–20, 1996.

[2] A. Efros and T. Leung, "Texture synthesis by non-parametric sampling," in *Proc. of IEEE ICCV'99*, Greece, pp. 1033–1038, Sep. 1999.

[3] A. Efros and W.T. Freeman, "Image Quilting for Texture Synthesis and Transfer," *Proceedings of SIGGRAPH '01*, Los Angeles, California, Aug. 2001.

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

[5] A.W. Fitzgibbon and A. Zisserman, "Automatic 3D model acquisition and generation of new images from video sequences," in *Proc. of European Signal Processing Conference*, Greece, pp. 1261–1269, 1998.

[6] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, pp. 721–741, 1984.

[7] A. Hertzmann, C.E. Jacobs, N. Oliver, B. Curless, D.H. Salesin, "Images Analogies," *Proceedings of SIGGRAPH '01*, Los Angeles, California, Aug. 2001.

[8] A. Hilton, "Reliable Surface Reconstruction from Multiple Range Images," in *Proc. of ECCV*, 1996.

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

[10] M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, J.'Ginsberg, J. Shade, and D. Fulk, "The Digital Michelangelo Project: 3D scanning of large statues," in *SIGGRAPH*, 2000.

[11] J. Oliensis, "Uniqueness in Shape From Shading," in the *Int. Journal of Computer Vision*, Vol. 6(2), pp. 75–104, 1991.

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

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

[14] V. Sequeira, K. Ng, E. Wolfart, J.G.M. Goncalves, D.C. Hogg, "Automated Reconstruction of 3D Models from Real Environments," in *ISPRS Journal of Photogrammetry and Remote Sensing* (Elsevier), Vol. 54, pp. 1–22, Feb. 1999.

[15] I. Stamos and P.K. Allen, "3D Model Construction using range and image data," in *CVPR 2000*, South Caroline, Jun. 2000.

[16] K. Storjohann, "Laser Range Camera Modeling," technical report ORNL/TM-11530, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 1990.

[17] L. Wei and M. Levoy, "Fast texture synthesis using tree-structured vector quantization," in *SIGGRAPH 2000*, pp. 479–488, Jul. 2000.