Image and Sparse Laser Fusion for Dense Scene Reconstruction

Alastair Harrison and Paul Newman.

Abstract This paper is concerned with reconstructing the metric geometry of a scene imaged with a single camera and a scanning laser. Our aim is to assign each image pixel with a range value using both image appearance and sparse laser data. We pose the problem as an optimisation of a cost function encapsulating a spatially varying smoothness cost and measurement compatibility. In particular we introduce a second order smoothness term. We derive cues for discontinuities in range from changes in image appearance and reflect this in the objective function. We show that our formulation distills down to solving a large linear system which can be solved swiftly using direct methods. Results are presented and analysed using synthetic cases to demonstrate salient behaviours and on real data to highlight real-world applicability.

1 Introduction and Motivation

This paper is about dense mapping of workspaces using common place cameras and scanning lasers. Cameras provide near instantaneous capture of the workspace's appearance (texture and colour) but, from a single view, little geometrical information. On the other hand, scanning lasers produce comparatively slow, sparse metric sampling and beyond reflectance, capture little of the scene's appearance. This motivates us to consider how we might fuse sparse laser data and images to infer a range for every pixel in the image, allowing us to reconstruct a 3D scene with all the texture, colour and appearance information captured in the original image. The heart of the problem is how to sensibly infer ranges for pixels which are not near any laser measurements without introducing intolerable distortions. Our method is general in that it is not tied to any particular 3D laser scanner mechanism or geometry. Note also that we aim to recover the dense geometry of a scene over scales which prohibit the use of other direct methods such as stereo unless a truly large baseline is used.

Alastair Harrison
University of Oxford, Oxford, OX1 3PJ, e-mail: arh@robots.ox.ac.uk

Paul Newman
University of Oxford, Oxford, OX1 3PJ, e-mail: pnewman@robots.ox.ac.uk

2 Related Work

The problem of inferring 3D surface models of a scene using laser or camera sensors has been studied extensively over many years (see, for example [1, 2, 3, 4]). However, limitations in hardware and a requirement for speedy data gathering in mobile robotics typically results either in high resolution optical images only allowing inference of very basic 3D geometry, or, alternatively, low resolution range images which often sample the scene too sparsely to allow for faithful reconstruction. Multiple view reconstruction provides an attractive alternative due to a near instantaneous gathering of dense 3D data leading to dense scene reconstructions from image data alone [5, 6]. Unfortunately, stereo reconstruction fidelity is limited in range by the baseline and the image resolution. This seriously impedes accurate reconstruction beyond a few meters from the camera. Another alternative can be found in the exploitation of the complementary nature of vision and range sensing. While optical images and range images represent different quantities, they share "similar second order statistics and scaling properties" [7].

Only a relatively small body of work exists on the inference of surfaces by fusing laser data and camera images. Usually, these techniques exploit the fact that edges in the optical image often correspond to discontinuities in depth, and that smooth surfaces tend to correspond to areas of similar colour and texture. In [8], depth values for pixels in an image are inferred using belief propagation in a Markov Random Field (MRF) framework. The technique requires that the supplied range measurements contain some high density areas from which to seed the solution, and is unable to assign depth values outside of those already in the measurements. The techniques described in [9], [10] and [7] are able to fuse the information from both sources to significantly improve the resolution of low quality range images. The method of [9] is particularly relevant to this work. It employs an MRF formulation with a first-order smoothness prior. The technique favours fronto-parallel surfaces, but does not suffer too greatly from this because the range measurements are sufficiently regular and dense, coming from a special range camera sensor. This 'pins' the estimates to lie near the true surface.

In contrast to [9] the method presented here is targeted at any combination of commonly available monocular camera and scanning laser. In particular, this requires inference of range measurements based on sparse, inhomogenous range data. In such cases, the fronto-parrallel tendency of inferred surfaces induced by only considering a first-order smoothness prior leads to increasingly inaccurate reconstructions. We address that issue by introducing a second-order smoothness prior while still framing the problem as a well-understood optimization of a linear system of equations.

3 Problem Formulation

In this section we shall show how a general description of the problem can be formulated in such a way that in the end, only the solution of a single linear system is required. We begin by introducing our notation.

We are given a u by v pixel image $\mathscr{I}$ and a $3D$ point cloud of k laser measurements $\mathscr{L} = \{l_1 \cdots l_k\}$. We shall use the notation $\mathbf{I}_i$ to represent the i^{th} pixel in a vectorised image (all pixels stacked in a single vector of length $N = u \times v$). For each $\mathbf{I}_i$ we associate a range x_i. Our task is to use both $\mathscr{I}$ and $\mathscr{L}$ to find a vector $\mathbf{x} = [x_1, x_2 \cdots x_N]^T$ - a range for every pixel in the image. We shall also refer to x_i as a "range node". Each point in $\mathscr{L}$ can be projected into $\mathscr{I}$ under a distortion correcting camera model and associated to the nearest pixel. Each laser point then yields a range measurement z_i tied to pixel $\mathbf{I}_i$. Note the laser measurements are sparse so not every pixel will have a range measurement — in fact very few will. We use the notation $i \in \mathscr{L}$ to imply the index variable i ranges over all pixels which have an associated range measurement.

We shall pose the problem as one of finding the optimal range vector $\mathbf{x}^*$ such that

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \{\lambda_1 \lambda_2 \Theta_s(\mathbf{x}, \mathbf{I}) + \lambda_1 (1 - \lambda_2) \Theta_c(\mathbf{x}, \mathbf{I}) + (1 - \lambda_1) \Theta_d(\mathbf{x}, \mathbf{z})\} \tag{1}$$

where $\Theta_s(\mathbf{x}, \mathbf{I})$ is a first order cost penalising depth discontinuities, $\Theta_c(\mathbf{x}, \mathbf{I})$ is a second order cost penalising curvature and $\Theta_d(\mathbf{x}, \mathbf{z})$ is a data cost penalising errors between inferred ranges and observed range measurements. The scalars $\lambda_1, \lambda_2 \in [0, 1]$ are weightings between the three terms. We shall now consider these terms in more detail.

3.0.1 Data Cost

The data cost is defined as a squared error between assigned range, x_i and measured range, z_i

$$\Theta_d(\mathbf{x}, \mathbf{z}) = \sum_{i \in \mathscr{L}} \sigma_i (x_i - z_i)^2 \tag{2}$$

$$= ||\mathbf{W}(\mathbf{x} - \mathbf{z})||^2 \tag{3}$$

where $\mathbf{W}$ is a diagonal matrix with entries

$$\mathbf{W}_{i,i} = \begin{cases} \sigma_i & \text{if } i \in \mathscr{L} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and σ_i is a measure of our confidence in measurement z_i.

3.0.2 Discontinuity Cost

As in [9], we use a depth smoothness or *first-order* prior of the form

$$\Theta_s(\mathbf{x}, \mathbf{I}) = \sum_i \sum_{j \in \mathscr{N}(i)} e_{i,j} (x_i - x_j)^2 \tag{5}$$

where $\mathscr{N}(i)$ are the horizontal and vertical neighbours of i. As edge strength between nodes we use an exponentiated L_2 norm of the difference in pixel appearance

$$e_{i,j} = \exp - \frac{||\mathbf{c}_i - \mathbf{c}_j||^2}{\sigma_d^2} \tag{6}$$

where $\mathbf{c}_i$ is the RGB colour vector of pixel i and σ_d is a tuning parameter (small σ_d increases sensitivity to changes in the image). Equation 5 may be written in matrix form as

$$\Theta_s(\mathbf{x}, \mathbf{I}) = ||\mathbf{S}\mathbf{x}||^2 \tag{7}$$

where each row of $\mathbf{S}$ represents a weighted average of a pair of adjacent range nodes.

3.0.3 Smoothess/Curvature Cost

In contrast to [9] we make the further assumption that in the absence of cues to the contrary, such as discontinuities in appearance, the gradient of surfaces varies smoothly. Under this *second order* smoothness assumption, given a neighbourhood $\mathcal{N}(i)$ of node x_i we may make a range prediction $\hat{x}_i$ as a linear combination of neighbouring ranges x_j for $j \in \mathcal{N}(i)$. This allows us to write simply

$$\hat{\mathbf{x}} = \mathbf{P}\mathbf{x} \tag{8}$$

where $\mathbf{P}$ is a suitably formed prediction matrix. We define curvature cost $\Theta_c(\mathbf{x}, \mathbf{I})$ in the form

$$\Theta_c(\mathbf{x}, \mathbf{I}) = ||\hat{\mathbf{x}} - \mathbf{x}||^2 \tag{9}$$

$$= ||(\mathbf{P} - \mathbf{1})\mathbf{x}||^2 \tag{10}$$

Here, $\mathbf{1}$ is the identity matrix. While details of how $\mathbf{P}$ is created will be postponed until Section 4 we may proceed by understanding this cost as the L_2 norm of the deviation of $\mathbf{x}$ from the prediction based on modeling surfaces as locally continuous and smooth.

3.1 Reduction to $Ax = b$

We may further expand Equation 3 to the form

$$\Theta_d(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\mathbf{T}\mathbf{W}^\mathbf{T}\mathbf{W}\mathbf{x} - 2\mathbf{z}^\mathbf{T}\mathbf{W}^\mathbf{T}\mathbf{W}\mathbf{x} + \mathbf{z}^\mathbf{T}\mathbf{W}^\mathbf{T}\mathbf{W}\mathbf{z} \tag{11}$$

and Equations 10 and 5 to

$$\Theta_c(\mathbf{x}, \mathbf{I}) = \mathbf{x}^\mathbf{T}\mathbf{R}^\mathbf{T}\mathbf{R}\mathbf{x}, \quad \Theta_s(\mathbf{x}, \mathbf{I}) = \mathbf{x}^\mathbf{T}\mathbf{S}^\mathbf{T}\mathbf{S}\mathbf{x} \tag{12}$$

where $\mathbf{R} = \mathbf{P} - \mathbf{1}$.

Substituting Equations 11, 12 into 1 and solving for $\mathbf{x}$ reduces the problem to

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{13}$$

with

$$\mathbf{b} = \mathbf{W}^{\mathbf{T}}\mathbf{W}\mathbf{z} \tag{14}$$

$$\mathbf{A} = \frac{\lambda_1\lambda_2\mathbf{R}^{\mathbf{T}}\mathbf{R} + \lambda_1(1-\lambda_2)\mathbf{S}^{\mathbf{T}}\mathbf{S} + (1-\lambda_1)\mathbf{W}^{\mathbf{T}}\mathbf{W}}{1-\lambda_1} \tag{15}$$

Equations 13 to 15 imply that all we need to do to perform the optimization is to solve a large sparse linear system.

4 Constructing The Prediction Matrix

In this section we detail how the prediction matrix **P** is created. For simplicity we show only 1D cases but it should be noted that **P** contains elements to penalise curvature in *both horizontal and vertical* directions.

We decompose **P** into a weighted sum of three prediction operators - extrapolation from left and right, and interpolation.

$$\mathbf{P} = \mathbf{W}_L\mathbf{P}_L + \mathbf{W}_M\mathbf{P}_M + \mathbf{W}_R\mathbf{P}_R \tag{16}$$

where subscripts *L*,*M*,*R* imply left-extrapolation, mean (interpolation) and right-extrapolation respectively. The **W**'s are suitably constructed weighting matrices derived from image appearance which we shall expand upon shortly in Section 4.1. The use of extrapolation and interpolation can be understood graphically with reference to Fig. 1 which shows a simplified 1D case.

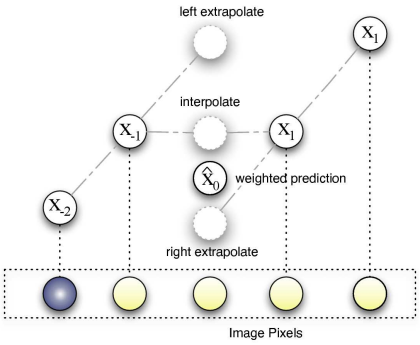


Fig. 1 Depth prediction via weighted interpolation and extrapolation in 1D. The predictions of the range x_0 by left and right extrapolation and interpolation are shown in faded grey. The discontinuity in the image shown at the bottom of the figure (each range node has a single pixel attached to it) causes the left extrapolation to be down-weighted — the image edge is a cue for a possible discontinuity in range between node x_{-1} and x_o. The final prediction, $\hat{x}_0$ is shown in the center.

4.1 Anticipating Depth Discontinuities from Image Cues

The image $\mathscr{I}$ can be used to provide cues about the behaviour of the surface we hope to reconstruct. Our basic assumption is one that has been used before [9] — sharp changes in range tend to appear as changes in appearance (edges) in an image. We have a range node for each pixel (see Equation 16) and its value can be predicted by a weighted sum of extrapolation and interpolation from its neighbours. We describe only the horizontal case for simplicity, but our method is applied in the vertical case too. For each node x_i the weighting is determined by the properties of pixel i and its neighbourhood. Broadly speaking, if a pixel is identical to its left and right neighbours then pure interpolation will occur. If however there is a discontinuity in pixel appearance then interpolation will be down weighted and either left or right extrapolation emphasised.

To explain how the weighting matrices $\mathbf{W}_{L,M,R}$ are created we shall consider the simple 1D case shown in Fig. 2. Interpolation is preferable to extrapolation. With this preference in mind and considering node x_0 in Fig. 2, we can write the importance weights of left / right extrapolation and interpolation as $w_{l,m,r}$

$$w_m = e_{(-1,0)}e_{(0,1)} \tag{17}$$

$$w_r = e_{(-2,-1)}e_{(-1,0)}(1-w_m) \tag{18}$$

$$w_l = e_{(2,1)}e_{(1,0)}(1-w_m) \tag{19}$$

with $e_{i,j}$ as defined in Equation 6. The above relationships can be understood by noting that if the pixel attached to range node x_0 is identical to its neighbours ($e_{(-1,0)}$ and $e_{(0,1)}$ are unity) then $w_m = 1$ and $w_r = w_l = 0$ - interpolation has 100% of the weighting. As the pixels $\mathbf{I}_{-1}$ and $\mathbf{I}_1$ become increasingly different, the left and right extrapolations receive more weight. In the limit, if two pixels are entirely different, the edge weight between them tends to zero and the attached range nodes will have no direct link between them. It does not make the two nodes independent - there may be other dependencies via long circuitous routes through other nodes. It does however mean that range discontinuities across this boundary are not penalised because the range prediction made by multiplication by $\mathbf{P}$ is based on an extrapolation from one side and not an interpolation across the discontinuity. This is a key point in this work.

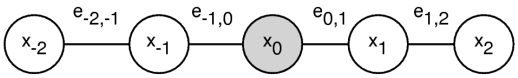
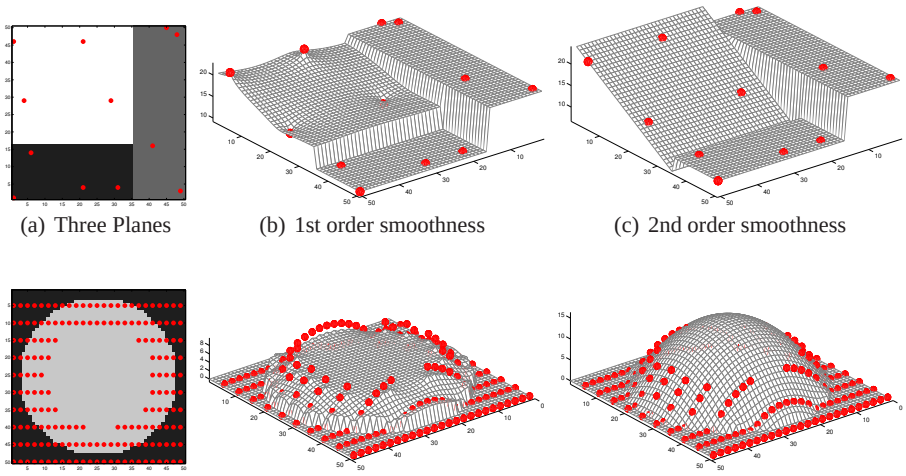


Fig. 2 A 1D chain of range nodes (a section of $\mathbf{x}$) and the edges between neighbours. Considering x_0, right extrapolation uses only nodes to the right and left extrapolation uses the two left hand nodes. Interpolation uses nodes x_{-1} and x_1. The edges between nodes are a function of the difference in pixel appearance between adjacent range nodes (each range node is associated with a single pixel in the image).

5 Results

Fig. 3 shows the results of processing two synthetic scenes. In this case the problem size is small with $\mathbf{x}$ having just 2500 elements (each element of $\mathbf{x}$ corresponds to a vertex in the mesh). With regard to the "three plane" case note how using just a few laser points in each distinct region of the image results in three distinct planes being generated in the reconstructed scene. The strong edges in the images prohibit information flow between planes. For the nodes at the very edge of a plane the extrapolation and interpolation weights have become such that the node is only influenced by (coupled to) other in-plane nodes. The 1st order method alone is unable to reconstruct the planes correctly as it tries to make all nodes have similar ranges.

In the case of the "dome" example note how while there is no range discontinuity there is a sharp discontinuity in surface gradient around the perimeter of the dome. Note also that the first order smoothness term is unable to reconstruct the curvature of the dome in the absence of laser measuremnents. In contrast, with a second order smoothness cost the curved shape of the dome is recovered well. This is an important result. The generated curved surface is the smoothest surface that can explain the existing measurements and minimise the bust in second order smoothness constraints implicit in $\mathbf{P}$.



(a) Three Planes (b) 1st order smoothness (c) 2nd order smoothness

(d) Dome Image (e) 1st order smoothness (f) 2nd order smoothness

Fig. 3 Synthetic data examples which highlight important aspects of our approach. Each node in the mesh represents a single range node projected out from an image pixel. The images for each of the two cases are shown on the left. In all figures sparse laser measurements are shown in red. Note how the discontinuities in the image appear as discontinuities in the reconstructed surfaces. First-order smoothness alone tends to make surfaces have the same depth value whereas second-order smoothness is able to correctly reproduce both planar and curved surfaces.

We now turn to processing some real data. We used a nodding SICK LMS200 laser scanner on a mobile robot to capture laser data. Images were captured by a camera

mounted above the laser with a wide angle lens. The image used in this case was 518 by 259 pixels resulting in some 134,162 range nodes and is shown in Fig. 4 with laser measurements projected into it. For scale, the target is approx 1.7m wide. The reconstructed model is shown alongside. Using second-order smoothness alone provides reasonable results, but tends to introduce 'rippling' artefacts around noisy measurements. A small amount of first-order smoothness is necessary to damp the oscillations. Fig. 5 shows points of interest in the reconstruction. We show an outdoor result of the same problem size in Fig. 6.

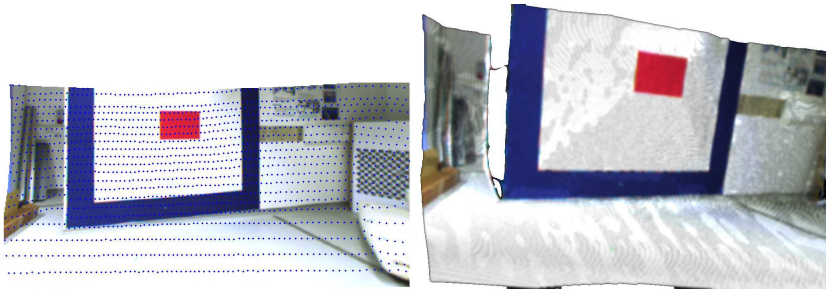


Fig. 4 Results from an indoor dataset. Image and laser measurements on the left, and the reconstructed model on the right.

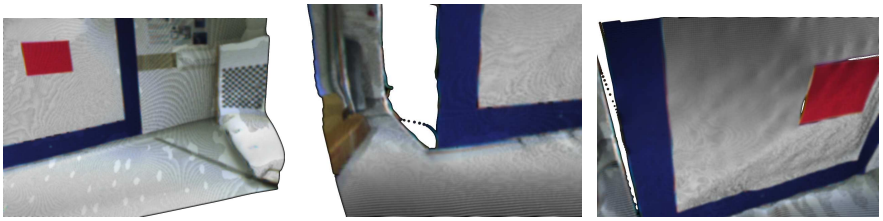


Fig. 5 Details of a reconstructed scene from Fig. 4. Note the detail of the smooth floor and inferred sharp range discontinuity between two walls.

The algorithm is implemented in Matlab and the linear solve is performed with Matlab's backslash operator (though there is no reason not to use another method such as Conjugate Gradient). The Three Planes case and the Dome case in Fig. 3, with 2,500 nodes both took 0.021 seconds to solve in a single iteration. For the real data case in Fig. 4 with 134,162 nodes, the algorithm took around 30 seconds on a 2Ghz dual core laptop.

We now present some numerical analysis of the performance of our approach. It is a hard task to obtain a ground truth geometry for the complete real scene. Instead of comparing pixel ranges to ground truth we compare them to laser measurements taken of the scene over a long period of time and which are not used in the optimisation. Concretely, we collect a very dense cloud of laser data at the scene and draw from that a small sparse test set with which we reconstruct the scene shown in Fig. 4.

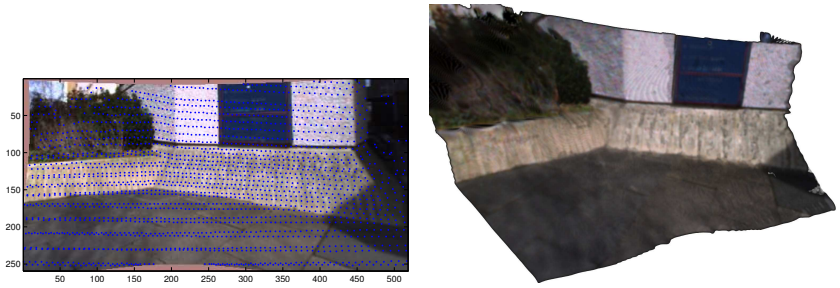
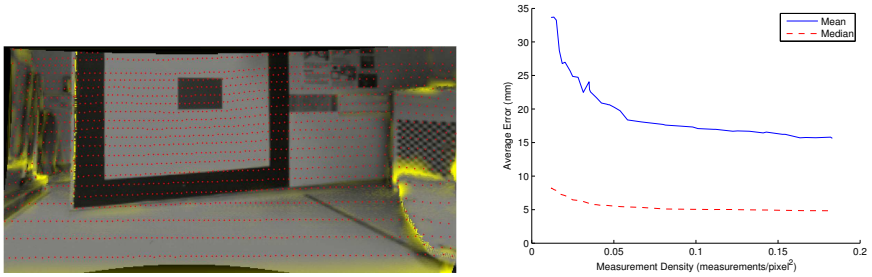


Fig. 6 Results from an outdoor dataset. On the left is the image with laser measurements overlaid. On the right is the reconstructed model.

The remaining laser data constitutes a dense hold out set, and for each unused laser measurement we can compare measured range to estimated range. Fig. 7(a) shows regions of the workspace which contain pixels with significant errors.



(a) (b)

Fig. 7 The left image shows a comparison of range estimates to ground truth laser data for the indoor case. Areas in yellow show deviation from ground truth, with higher intensity representing larger errors. Laser measurements are shown in red. The graph shows average error of the estimate relative to the mean density of range measurements, when compared to laser measurements in the hold out set. The laser has a precision of 15mm.

It is also instructive to consider how the accuracy of our approach depends on the density of laser measurements. Fig. 7(b) shows how the statistics (mean and median) of the pixel range errors change as a function of measurement density. Note that as expected, as measurement density increases the precision tends to that of the laser itself around 15mm. The results given in Figs. 4 and 6 are operating in the 0.01 measurements/pixel^2 region.

6 Conclusion

This paper has introduced a novel technique for fusing sparse laser data and images to enable a dense 3D scene reconstruction. Above and beyond existing prior work this

technique uses a second order smoothness term which allows it to extrapolate both planar and curved surfaces. The problem is formulated as the solution of a sparse linear system, which allows the use of fast optimization techniques. The technique was applied to both illustrative synthetic cases as well as real data recorded in indoor and outdoor scenes containing challenging geometry.

The qualitative and quantitative results presented here suggest that our system provides 3D reconstructions of reasonable quality. Nevertheless, there is room for improvement. In particular we must consider how we can increase robustness to erroneous laser measurements (away from image edges) and how we might fuse multiple scenes in a principled way. The flip side of this problem is handling bona-fide discontinuities in range when there is no change in image appearance and vice versa.

Acknowledgment

This work described here has been supported by the UK EPSRC (CNA and Platform Grant EP/D037077/1), the Office of Naval Reasearch and the European commission under grant agreement number FP7-231888-EUROPA. The authors would also like to thank Ingmar Posner for his invaluable contributions.

References

1. D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in *In CVPR*, 2006, pp. 2137–2144.
2. A. Saxena, S. H. Chung, and A. Y. Ng, "3-d depth reconstruction from a single still image," *Int. J. Comput. Vision*, vol. 76, no. 1, pp. 53–69, 2008.
3. H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, "Surface reconstruction from unorganized points," in *SIGGRAPH '92: Proceedings of the 19th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1992, pp. 71–78.
4. Y. Ohtake, A. Belyaev, M. Alexa, G. Turk, and H.-P. Seidel, "Multi-level partition of unity implicits," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 463–470, 2003.
5. D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.
6. R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.
7. Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on.*, vol. 0, pp. 1–8, 2007.
8. L. A. Torres-Méndez and G. Dudek, "Statistics of visual and partial depth data for mobile robot environment modeling," in *MICAI*, 2006, pp. 715–725.
9. J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2005.
10. H. Andreasson, R. Triebel, and A. Lilienthal, "Vision-based interpolation of 3D laser scans," in *Proc. International Conference on Autonomous Robots and Agents (ICARA)*, 2006.