Experience With Bayesian Image Based Surface Modeling

John C. Stutz

NASA Ames Research Center, Moffett Field, CA 94035, USA

Abstract. Bayesian surface modeling from images requires modeling both the surface and the image generation process, in order to optimize the models by comparing actual and generated images. Thus it differs greatly, both conceptually and in computational difficulty, from conventional stereo surface recovery techniques. But it offers the possibility of using any number of images, taken under quite different conditions, and by different instruments that provide independent and often complementary information, to generate a single surface model that fuses all available information.

I describe an implemented system, with a brief introduction to the underlying mathematical models and the compromises made for computational efficiency. I describe successes and failures achieved on actual imagery, where we went wrong and what we did right, and how our approach could be improved. Lastly I discuss how the same approach can be extended to distinct types of instruments, to achieve true sensor fusion.

Joint work with Peter Cheeseman, Frank Kuehnel, Andre Jalobeanu, Esfandiar Bandari, Doran Tal, Eudo von Toussaint, Robin Morris, Vadim Smelyanskiy, and David Maluf.

Keywords: Bayesian inference, vision, generative modeling, remote sensing, data fusion
PACS: 02.50.Tt, 07.05.Tp, 07.05.Pj, 07.05.Kf, 02.60.Pn

SuperRes is a system for inferring super resolved 3-D surface models from multiple images taken from diverse viewpoints and under diverse lighting. The project's primary objective was to show that, from such an image set, it is possible to accurately infer 3-D surface properties at a scale smaller than the projected pixel footprints. This was achieved, albeit to a limited extent, using CMOS camera images of a known physical surface. The project's significance lies in having successfully demonstrated that the combination of generative modeling driven by Bayesian Inference can be made to work in this context, and in identifying the problem areas that must be addressed to make such systems work well.

The SuperRes project was conceived and led by Peter Cheeseman, with initial development by Robin Morris. Robin Morris, Vadim Smelyanskiy and David Maluf designed and implemented the current program, carried out the initial tests on synthetic data, and wrote most of the descriptive papers. Frank Kuehnel and Udo von Toussaint tracked down and corrected some particularly nasty bugs that first surfaced with real images, while also elucidating several fundamental aspects of the representation. Working in parallel, Andre Jalobeanu developed a renderer capable of correctly handling occlusions, which has yet to be incorporated into the system. The author's contribution was largely limited to making the minimal extensions needed to represent a real camera, collecting suitable data, and devising the protocols needed to demonstrate super-resolved surface inference.

Full details of the underlying models, the rendering and derivative computations, and the optimization methods, are given in [1, 2, 3, 4, 5, 6] with [7] the best detailed exposition. This paper provides only an overview of the modeling and inference process, and concentrates on our efforts to apply the initial system to real data, the modifications needed to accomplish this, some results, and the lessons learned.

SuperRes System

Figure (1) illustrates the basic modeling problem for inference from imagery. From processed input images, we seek to infer as much as possible about the imaged surface. In addition to the unknown surface, there are potential uncertainties in the illumination, reflection, transmission, camera poses, camera projection, sensor response, digitization, and post processing. Depending on the specific application, some of these may be ignored, and others may be fixed by calibration. The remainder must be modeled, inferred, and preferably marginalized out the final result.

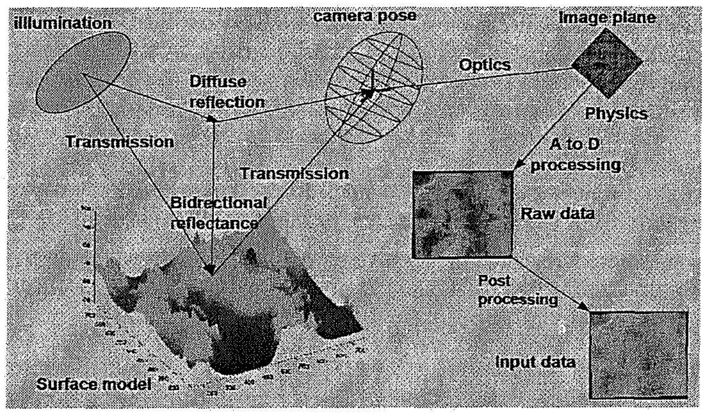

FIGURE 1. Overview of the modeling problem for remote sensing imagery.

Initially intended for extraterrestrial remote sensing, SuperRes modeled the remote surface as a triangulated mesh over a static rectangular grid of height field vertices $\vec{z}$. Monochrome albedos $\vec{\rho}$ were assigned to the vertices, and triangle albedos averaged over the vertex albedos. We initially assumed that viewing conditions precluded both occlusions and shadows. Illumination was by a single infinity distant point source (sun), supplemented by uniform ambient light. Reflectance was taken to be Lambertian. Atmospheric effects were assumed to be negligible. Images were described by 6-DoF camera pose and specific illumination. Cameras were described by focal length, and sensor array dimensions. Purely geometric perspective projection was assumed, with the camera optical axis centered on and perpendicular to the sensor array. Sensor array elements were assumed to tile the image plane, and have uniform linear response to incident light. No provision was made for A/D conversion or post processing.

Height field vertices were then projected to the image plane and the precise overlap areas between projected triangles and pixels were determined. Thus the triangle energies radiated through the camera lens were allocated to the sensor elements in proportion to the intersection areas. Summing the triangle contributions to each pixel then gives the rendered image. In essence, we implemented a textbook model of a chip based imager.

The paired vectors of heights and albedos $\mathbf{u} = [\vec{z}\,\vec{\rho}\,]$ form our full surface model. Their prior distribution is assumed to be Gaussian:

$$p(\vec{z},\vec{\rho}\,) \propto Exp\left(-\frac{1}{2}\mathbf{u}\Sigma^{-1}\mathbf{u}^T\right), \qquad \Sigma^{-1} = \begin{bmatrix} \hat{Q}/\sigma_h^2 & 0 \\ 0 & \hat{Q}/\sigma_\rho^2 \end{bmatrix}, \tag{1}$$

where the inverse covariance matrix Σ is constructed to enforce a smoothing constraint on local variations of heights and albedos. We penalize the integral over the surface of the curvature factor $c(x,y) = z_{xx}^2 + z_{yy}^2 + 2z_{xy}^2$, and similarly for albedos. We approximate the partial derivatives in $c(x,y)$ using finite differences of the height (albedo) values. The two hyper-parameters σ_h and σ_ρ in equation (1) control the expected values of the surface-averaged curvatures for heights and albedos. Since albedos are conceptually limited to the 0-1 range, we actually apply this prior to the Log-Odds transformed albedos.

Our likelihood assumes that the differences between observed and synthesized pixels are conditionally independent, with a zero mean Gaussian distribution. The negative log-posterior is then:

$$L(\vec{z},\vec{\rho}\,) \propto \frac{\Sigma_{f,p}(I_{fp} - \hat{I}_{fp}(\vec{z},\vec{\rho}\,))^2}{\sigma_e^2} + \mathbf{x}\Sigma^{-1}\mathbf{x}^T, \tag{2}$$

where $\hat{I}_{fp}(\vec{z},\vec{\rho}\,)$ denotes the pixel intensities in the image f synthesized from the model, σ_e^2 is the noise variance and the summation is over the pixels (p) and over all images (f) used for the inference. Vector $\mathbf{x} = \mathbf{u} - \mathbf{u}_0$ is a deviation from a current estimate $\mathbf{u}_0$.

In the initially assumed absence of shadows and occlusions the log-posterior is in general unimodal and gradient methods can be applied for minimizing $L(\vec{z},\vec{\rho}\,)$. We linearize $\hat{I}(\vec{z},\vec{\rho}\,)$ about the current estimate, $\vec{z}_0, \vec{\rho}_0$

$$\hat{I}(\vec{z},\vec{\rho}\,) = \hat{I}(\vec{z}_0,\vec{\rho}_0) + \mathbf{D}\mathbf{x}, \quad \mathbf{D} \equiv \left\{ \frac{\partial \hat{I}_{fp}}{\partial z_i}, \frac{\partial \hat{I}_{fp}}{\partial \rho_i'} \right\} \tag{3}$$

where $\mathbf{D}$ is the matrix of derivatives evaluated at z_0, ρ_0. Then the minimization of $L(\vec{z},\vec{\rho}\,)$ is replaced by minimization of the quadratic form:

$$L' = \frac{1}{2}\mathbf{x}\hat{A}\mathbf{x} - \mathbf{b}\mathbf{x}, \tag{4}$$

$$\mathbf{x} \equiv \mathbf{u} - \mathbf{u}_0, \quad \hat{A} = \Sigma^{-1} + \frac{\mathbf{D}\mathbf{D}^T}{\sigma_e^2}, \quad \mathbf{b} = \frac{(I - \hat{I}(\vec{z}_0,\vec{\rho}_0))}{\sigma_e^2}\mathbf{D}.$$

Here $\hat{A}$ is the Hessian matrix of the quadratic form and vector $\mathbf{b}$ is the gradient of the likelihood L computed at the current estimate. We search for the minimum in $\mathbf{x}$ using a conjugate-gradient method. At the minimum we update the current estimate, $\mathbf{u}_1 = \mathbf{u}_0 + \mathbf{x}$, recompute $\hat{I}$ and $\mathbf{D}$, and repeat the minimization procedure iteratively until the current estimate $\mathbf{u}_k$ approaches the minimum of $L(\vec{z},\vec{\rho}\,)$.

Thus finding the MAP estimate requires that we can render the image and compute the derivatives w.r.t. the surface model parameters. While generating $\hat{I}$ via the modeling approach is computationally expensive, we can compute $\mathbf{D}$ at the same time for little additional effort. It helps that **D** is sparse, and becomes extremely so when triangle area is less than pixel footprint area.

The same approach is applied to the inference of camera and lighting parameters. Here the small number of parameters, of order 10 per image, is balanced by the fact that each influences essentially all of the image's pixels, giving a dense **D** matrix. When both surface and image parameters must be inferred, we alternate between the two, starting with a coarse surface mesh and successively refining it.

SuperRes Results

Early end to end testing, using synthetic data prepared with the camera model, gave remarkably satisfying results [1, 2, 7]. A digital surface model was constructed from a Geological Survey digital elevation model and textured with albedos from a corresponding Landsat image. With 16 low-resolution synthetic images, known camera positions and lighting, and inference starting from a flat medium gray plain, the inferred RMS height and albedo errors were of order 10^{-4} of the respective parameter ranges, on a surface modeled to half the pixel footprint spacing. Similar results were obtained on the task of inferring camera pose [8, 4].

However these excellent results merely proved that the inference system was working, and nothing about the ability of our camera model to match any specific real camera. Initial tests with scanned film camera and CCD camera images were complete failures. Switching to a CMOS camera, with its intrinsically cleaner response, gave no improvement. The problems were eventually traced to a combination of conceptual and implementation errors in the camera model.

Camera model implementation errors were of two types. The first were outright programming errors that were eventually tracked down and corrected. The second was more subtle, the assumption that the rules of arithmetic translate precisely into floating point arithmetic. This is only true for numbers of comparable magnitude. In our line and area intersection calculations we were caught out in assuming that the summed area of projected triangle intersections with a pixel must equal the pixel area. They did not, quite, and the resulting numerical noise was significant, and had to be explicitly compensated.

Camera model conceptual errors were much more serious. These were also dual. First, the camera model made no allowance for the optical misalignments and distortions that are inevitable in real cameras. Second, there was essentially no attempt to model the analog to digital conversion and output conditioning that convert incident light to a digital image. In essence, our textbook model was too simple to adequately describe real images.

Despite these errors, the basic approach proved sufficiently robust to give useful results after minimal modification. Adding off center projection and a linear output scaling proved sufficient to allow inference of reasonably shaped surfaces. Optical distortions were compensated by using Bouguet's Camera Calibration Toolbox [9] to generate undistorted projective images. In retrospect, this last was a mistake. We should have added the distortions to our camera model, even if we did not seek to estimate the parameters. These "undistorted" images are generated by resampling, which systematically distorts the pixels' numerical values.

Having obtained encouraging qualitative results from our CMOS images, we sought to quantify them. To this end, we had a 0.5m square surface model machined from a 300x300 point DEM. This has been sub-sampled from a USGS DEM, smoothed and vertically exaggerated by 5x, to get a semi-natural topographic surface with about 43mm height range. The source region includes several mountain ranges and level basins. This surface model was mounted on a 1.1m square checkerboard. When exposures included the full checkerboard, Bouguet's Calibration Toolbox [9] could determine camera pose and internal parameters to fair precision. A simple sundial gave reasonable initial estimates of sun direction. The whole was of a size that permitted it being moved outdoors for exposures under truly directional sun light and reasonably uniform ambient light. The only real problem was that camera positions were limited to within about 20 degrees of nadir, by the exaggerated topography, under our assumed absence of shadows and occlusions. Nor could we obtain images from near the sun direction, due to camera or tripod shadows on the model.

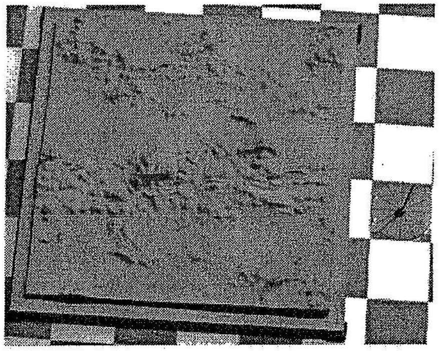

FIGURE 2. One of the 16 CMOS images used in our constant albedo experiment, showing the checkerboard and the sun dial used for sun calibration.

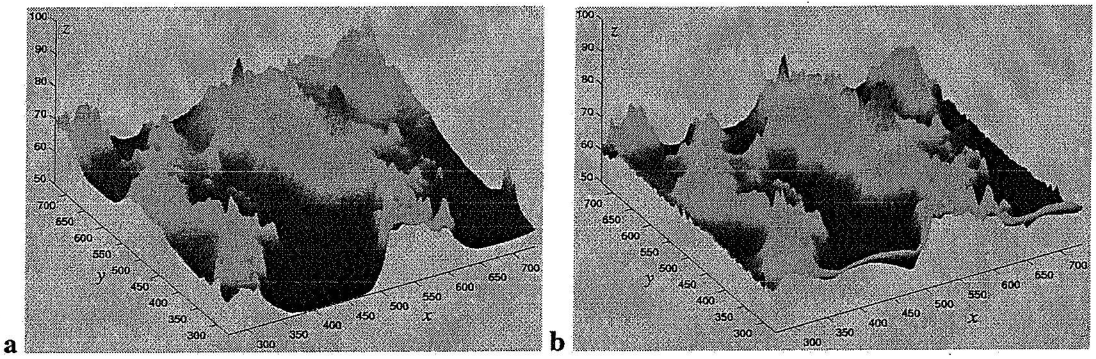

FIGURE 3. SuperRes results for the monochrome model: height field showing original topography(**a**) and estimated geometry model(**b**). The units are millimeters. Started from a flat surface at ~ 65mm elevation, using 5-step multi-grid.

Initial experiments were made with the surface model painted a uniform matte gray. Figure(2) shows a typical image, and figures(3, 4) compare the known and inferred surfaces obtained using 16 such images. This was achieved with a 5 step multigrid inference, working from 37x37 to 577x577 vertices, to avoid local minima. the starting point was a uniform gray level surface at approximate average elevation. Camera and

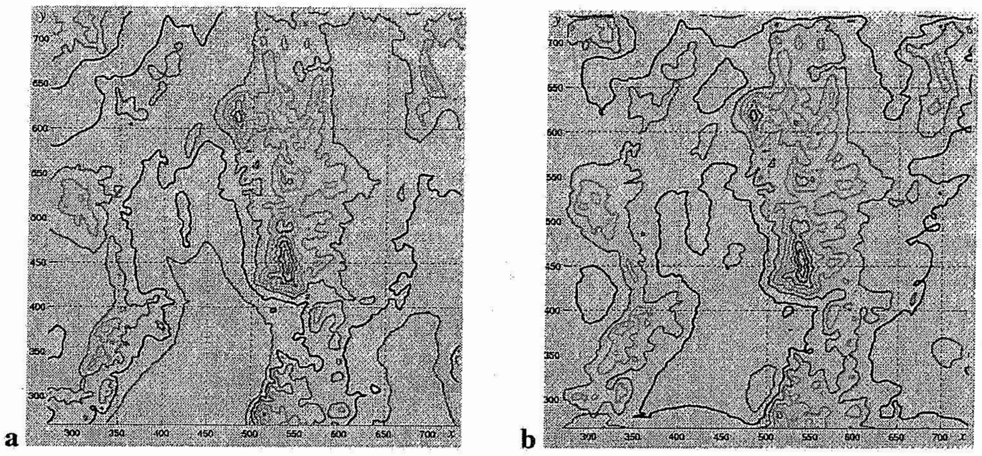

FIGURE 4. SuperRes results for the monochrome model: contour plot showing original topography(**a**) and estimated geometry model(**b**). The contours at 5mm intervals. Started from a flat surface at ~ 65mm elevation, using 5-step multi-grid.

lighting parameters were re-estimated prior to the last step, which vastly improves the surface estimates per [2]. The inference converged to an estimate that is close to the original model, with a maximum error less than 15mm, with a 2m distance between camera and model.

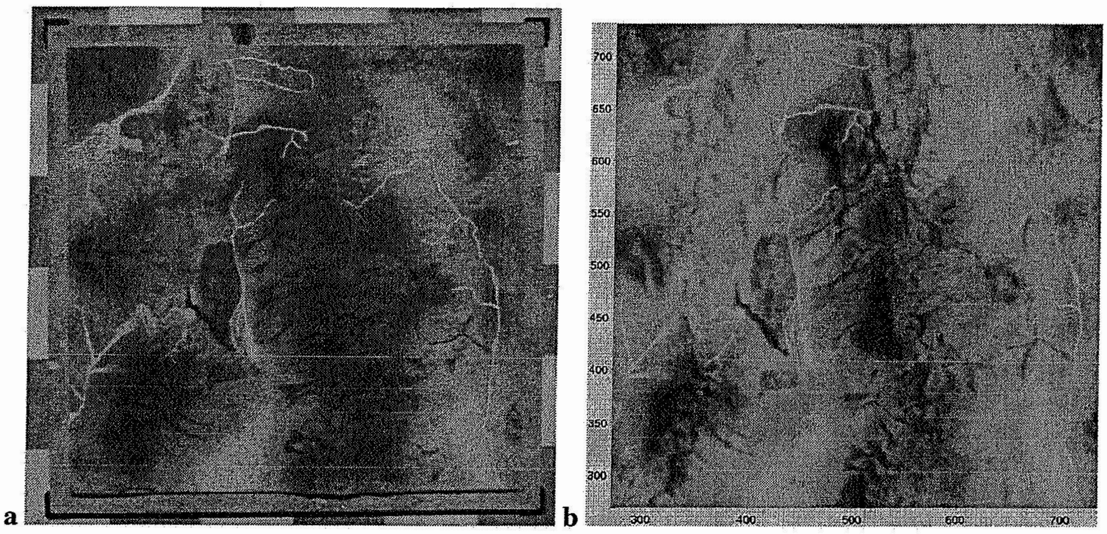

FIGURE 5. SuperRes with albedo variations: an original image(**a**) and a Matlab plot of the albedo field(**b**) inferred from a uniform 0.5 albedo and flat surface.

We made extensive experiments with the same physical model after adding albedo patterns(Fig. 5), experimenting with a number of variations of the basic surface inference / camera inference multigrid protocols. The albedo and height fields shown in figures 5 and 6 are typical of our better results. The inferred surface geometry shows RMS errors between 1 and 2mm, and maximum errors usually less than 10mm, which is better than in the constant albedo case. The inferred albedos, while visually acceptable, cannot be quantified, since we lack quantitative ground truth.

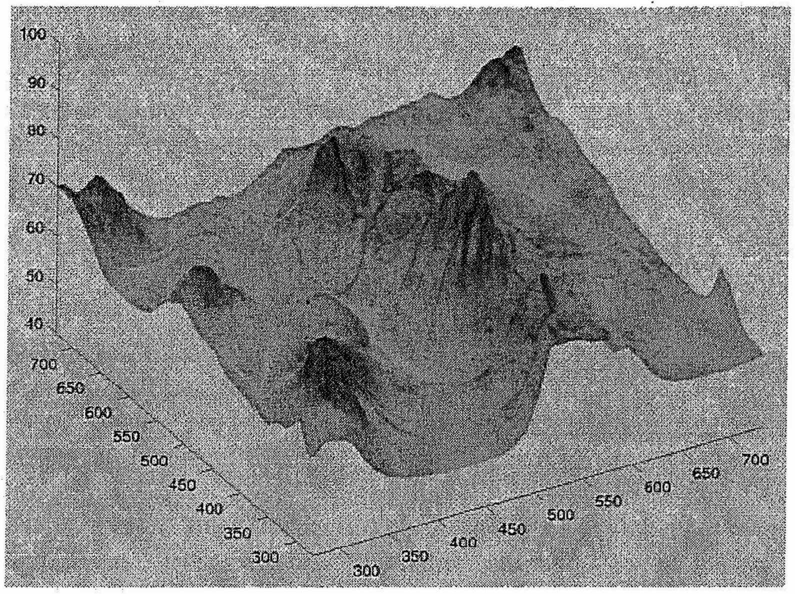

FIGURE 6. SuperRes inferred model with albedo variation. Peaks are somewhat flattened. Inference started with a flat surface and uniform albedo, achieving RMS height errors of <2mm, 10mm max error, with cameras 2m distant.

The albedo texture helps to recover the topography. However there is clearly some cross talk between albedos and geometry: slopes are influenced by the rate that albedo changes. Thus the linear albedo features tend generate channels or ridges, depending on contrast, in what should be level surfaces. This is believed due to our use of the curvature term in our surface priors. There are several possibilities for alleviating this problem, but none have yet been investigated in detail.

Sixteen images proved sufficient to recover a good surface, starting from a uniform albedo plain surface at the known mean elevation. These images were well distributed w.r.t camera positions and rotations and sun directions, with camera poses known to ~1% of distance, and sun direction to ~1 degree. Starting from a 37x37 height grid with uniform albedo, most details of the 289x289 known height grid are recovered with only 4 images. Most distortions can be attributed to cross-talk between the albedo and height fields where true albedos are rapidly changing.

Potential Extensions

There is nothing about this approach to surface modeling that limits it to a single camera. One could use several cameras, and simply adjust the parameters for each. One could even used quite distinct types of cameras, by providing appropriate camera models. Nor is there any intrinsic limitation to visual data. So long as an instrument records radiation that has interacted with the surface, and one can model both that interaction and the instrument, then this approach can be used to make inferences about the surface. Most significantly, when two such instruments record interactions with the same surface property, then their data can be jointly used to make common inference about that property, an example of model based data fusion.

Lidar provides a obvious complement to remote visual data. While vision gives fine angular discrimination, depths must be inferred. Lidar gives fine depth discrimination at relatively coarse angular spacing. Combined, the two should compensate for each other's deficiencies. Morris [3] has experimented with this. Using very simple synthetic lidar data to supplement synthetic visual data, he was able to recover excellent surface shape and albedo using only two images. Thus the lidar data greatly reduced the number of images needed to achieve comparable results with pure visual data.

Lessons Learned

First, inference is only as good as good as the model. We initially concentrated on achieving reliable and efficient inference. When we got it, we found that we had only solved half of our problem, since our camera model was too simple to match any real instrument. The data generating model must capture all significant factors that influence the data values.

Second, know your data. Second party data that has been subject to undocumented processing is particularly suspect, as we found to our great frustration, in atempting to use Mars Rover imagery.

Third, be wary of numerical results. Floating point is a sparse representation of the real numbers, calculations only approximate algebra, and approximations are applied at every step of a calculation. With large scale calculations, it can be folly to attempt to save time and space by shorting the precision of calculations.

REFERENCES

1. R. Morris, P. Cheeseman, V. Smelyanskiy, and D. Maluf, *Proc. of the IEEE Workshop on Higher Order Statistics*, pp. 140–143 (1999).
2. V. Smelyanskiy, P. Cheeseman, D. Maluf, and R. Morris, *Proc. of CVPR* (2000).
3. R. Morris, U. von Toussaint, and P. Cheeseman, *Proc. of the ISPRS Workshop on Land Surface Mapping and Characterization Using Laser Altimetry* (2001).
4. V. Smelyansky, R. Morris, F. Kuehnel, D. Maluf, and P. Cheeseman, *Proc. of ECCV* (2002).
5. A. Jalobeanu, "Bayesian Vision for Shape Recovery," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by R. Fischer, R. Preuss, and U. von Toussaint, AIP Conference Proceedings 735, American Institute of Physics, Melville, NY, 2004, pp. 143–152.
6. F. . Kuehnel, "Robust Bayesian estimation of nonlinear parameters on SE(3) Lie group," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by R. Fischer, R. Preuss, and U. von Toussaint, AIP Conference Proceedings 735, American Institute of Physics, Melville, NY, 2004, pp. 176–186.
7. V. Smelyansky, R. Morris, D. Maluf, and P. Cheeseman, (almost) featureless stereo - calibration and dense 3d reconstruction using whole image operations, Tech. Rep. TR 01.26, Research Insititute for Advanced Computer Science, NASA Ames Research Center, Moffett Field, CA 94035 (2001).
8. R. d. Morris, V. N. Smelyanskiy, and P. C. Cheeseman, "Matching Images to Models - Camera Calibration for 3-D Surface Reconstruction," in *Proceedings of EMMCVPR*, EMMCVPR, Springer, 2001, vol. Springr LNCS V.2134.
9. J.-Y. Bouguet, Camera calibration toolbox, `http://www.vision.caltech.edu/bouguet/calib_doc/` (n.d.).