Reconstructing Relief Surfaces

George Vogiatzis[1] Philip Torr[2] Steven M. Seitz[3] Roberto Cipolla[1]

[1] Dept. of Engineering, University of Cambridge, Cambridge,CB2 1PZ, UK

[2] Dept. of Computing, Oxford Brookes University, Wheatley, Oxford OX33 1HX, UK

[3] Dept. of Computer Science and Engineering

University of Washington Box 352350, Seattle, Washington 98195-2350, USA

gv215@eng.cam.ac.uk philiptorr@brookes.ac.uk

seitz@cs.washington.edu cipolla@eng.cam.ac.uk

Abstract

*This paper generalizes Markov Random Field (*MRF*) stereo methods to the generation of surface relief (height) fields rather than disparity or depth maps. This generalization enables the reconstruction of complete object models using the same algorithms that have been previously used to compute depth maps in binocular stereo. In contrast to traditional dense stereo where the parametrization is image based, here we advocate a parametrization by a height field over any base surface. In practice, the base surface is a coarse approximation to the true geometry, e.g., a bounding box, visual hull or triangulation of sparse correspondences, and is assigned or computed using other means. A dense set of sample points is defined on the base surface, each with a fixed normal direction and unknown height value. The estimation of heights for the sample points is achieved by a belief propagation technique. Our method provides a viewpoint independent smoothness constraint, a more compact parametrization and explicit handling of occlusions. We present experimental results on real scenes as well as a quantitative evaluation on an artificial scene.*

1 Introduction

Inferring the dense 3D geometry of a scene from a set of photographic images is a computer vision problem that has been extensively studied. Work in this area can be roughly divided into two classes: (1) techniques for computing depth maps (image-based parameterization), and (2) volumetric methods for computing more complete object models.

In the first class, *image based parameterization* of shape, a reference image is selected and a disparity or depth value is assigned to each of its pixels using a combination of image correlation and regularization. Scharstein and Szeliski provide an excellent review for image based methods [18]. These problems are often formulated as minimizations of Markov Random Field (MRF) energy functions providing a clean and computationally-tractable formulation, for which good approximate solutions exist [11, 12, 17, 20]. However, a key limitation of these solutions is that they can only represent depth maps with a unique disparity per pixel, i.e. depth is a funtion of image point. Capturing complete objects in this manner requires further processing to merge multiple depth maps [15], a complicated and error-prone procedure. A second limitation is that the smoothness term imposed by the MRF is viewpoint dependent, in that if a different view was chosen as the reference image the results could be quite different.

The second class of techniques uses a *volumetric parameterization* of shape. In this class are well-known techniques like Space Carving [13] and level-set stereo [5]. There are also hybrid approaches that optimize a continuous functional via a discrete quantisation [16]. While these methods are known to produce high quality reconstructions, running on high resolution 3D grids is very computationally and memory intensive. Furthermore their convergence properties in the presence of noise are not well understood, in comparison with MRF techniques, for which strong convergence results are known. For Space Carving in particular, there is also no simple way to impose surface smoothness constraints.

In principle MRF stereo methods could be extended to multiple views. The problem is that reasoning about occlusions within the MRF framework is not straightforward because of global interactions between points in space (see [12] for an insightful but costly solution for the case of multi-view depth-map reconstruction). In this paper, we propose extending MRF techniques to the multi-view stereo domain by recovering a general *relief surface*, instead of a depth map. We assume that a coarse *base surface* is given as input. In practice this can be obtained by hand, by shape-from-silhouette techniques or triangulating sparse image correspondences. On this base surface sample points are uniformly and densely defined, and a belief propagation algorithm is used to obtain the optimal height above each sample point through which the relief surface passes. The benefits of our approach are as follows:

1. General surfaces and objects can be fully represented and computed as a single relief surface.
2. Optimisation is computationally tractable, using existing MRF solvers.
3. Occlusions are approximately modelled.
4. The representation and smoothness constraint is image and viewpoint independent.

1.1 Related Work

Our work is inspired by displaced surface modelling methods in the computer graphics community, in particular the recent work of Lee et al. [14], who define a displacement map over subdivision surfaces, and describe a technique for computing such a representation from an input mesh. An advantage of this and similar techniques is that they enable the representation of finely detailed geometry using a simple base mesh.

We also build on work in the vision community on *plane-plus-parallax* [2], *model-based stereo* [4], and *sprites with depth* [19]. All of these techniques provide means for representing planes in the scene with associated height fields. Our work can be interpreted as a generalization of plane-plus-parallax to a surface-plus-height formulation.

Previous mesh-based multi-view stereo techniques operate by iteratively evolving an initial mesh until it best fits a set of images [10, 23], or depth maps [8]. Representing finely detailed geometry is difficult for such methods due to the need to manage large and complex meshes. In contrast we assume a fixed base surface and solve only for a height field providing a much simpler way of representing surface detail. We also use a more stable estimation problem with good convergence properties. Ultimately, a hybrid approach that combines surface evolution and height field estimation could offer the best of both worlds and is an interesting topic of future work.

2 Model

The theory of Markov random fields yields an efficient and powerful framework for specifying complex spatial interactions between a number of discrete random variables $h_1,\ldots,h_M$, usually called *sites*. Each site can take one of a number of values or *labels* $H_1,\ldots,H_L$. The first ingredient of the model is a labelling cost function $C_k(h_k)$ that measures how much a site is in agreement with being assigned a particular label. The second ingredient is the interaction between sites, which, in a pairwise MRF such as the one considered in this paper, is modelled through a symmetric neighbourhood relation $\mathscr{N}$ as well as a compatibility cost term $C_{kl}(h_k,h_l)$ defined over neighbouring sites. This cost term measures how compatible the assignment of any two neighbouring labels is. The cost of cliques (fully connected subgraphs) with more than two nodes is set to zero. With these energy functions defined, the joint probability of the MRF is:

$$Pr(h_1,\ldots,h_M) = \tfrac{1}{Z}\exp\left(-\textstyle\sum_{k=1}^{M} C_k(h_k) - \sum_{(k,l)\in\mathscr{N}} C_{kl}(h_k,h_l)\right) \tag{1}$$

where Z is a constant.

To bring multi-view stereo into this framework a set of 3D sample points $\mathbf{X}_1,\mathbf{X}_2,\ldots\mathbf{X}_M$ is defined on a *base surface*. The neighbourhood relation $\mathscr{N}$ was obtained by thresholding the Euclidean distance between sample points. At each sample point $\mathbf{X}_k$, the unit normal to the base surface at that point, $\mathbf{n}_k$ is computed. The sites of the MRF correspond to height values $h_1,\ldots,h_M$ measured from the sample points $\mathbf{X}_1,\mathbf{X}_2,\ldots\mathbf{X}_M$ along the normals $\mathbf{n}_1,\mathbf{n}_2,\ldots,\mathbf{n}_M$ (see fig. 1 left). The labels $H_1,\ldots,H_L$ are a set of possible height values that variables h_k can take. If the kth site is assigned label h_k then the *relief surface* passes through 3D point $\mathbf{X}_k + h_k\mathbf{n}_k$. To deal with the problem of occlusion, the base surface has to *contain* the relief surface for reasons that will be explained in the next section. Hence if the positive normal direction is defined to be towards the interior of the volume, only positive (inward) heights need be considered. The labelling cost is related to the photo-consistency [13] of the 3D point $\mathbf{X}_k + h_k\mathbf{n}_k$ while the compatibility cost forces neighboring sites to be labelled with 'compatible' heights. The following sections examine these two cost functions in more detail.

2.1 Labelling cost

The data are n images of the scene $I_1,\ldots,I_N$, with known intrinsic and extrinsic camera parameters. We will be denoting by $I_k(\mathbf{X})$ the intensity of the pixel onto which the 3D point $\mathbf{X}$ is perspectively projected by the camera that captured image I_k. As mentioned, labelling a site with a height value corresponds to a point in space through which the relief surface passes. Let that point be $\mathbf{X}_k + h_k\mathbf{n}_k$ and let the intensities of the pixels to which it projects be $\mathtt{i}_1(h_k) = I_1(\mathbf{X}_k + h_k\mathbf{n}_k), ..., \mathtt{i}_N(h_k) = I_N(\mathbf{X}_k + h_k\mathbf{n}_k)$. If the point is part of the true scene surface these intensities should be consistent. Let $\rho\left\{\mathtt{i}_1,\mathtt{i}_2,...,\mathtt{i}_n\right\}$ be some measure of consistency of the intensities. In experiments presented here this was set to the standard deviation of the intensities (which corresponds to the Lambertian reflectance model) but other measures could be used instead [9, 22]. Then

$$C_k(h_k) = w_1\rho\left\{\mathtt{i}_1(h_k),\ldots,\mathtt{i}_N(h_k)\right\} \tag{2}$$

is defined as a measure of the consistency of the assignment of height h_k to sample point $\mathbf{X}_k$ for some weight parameter w_1. This however does not take occlusion into account and

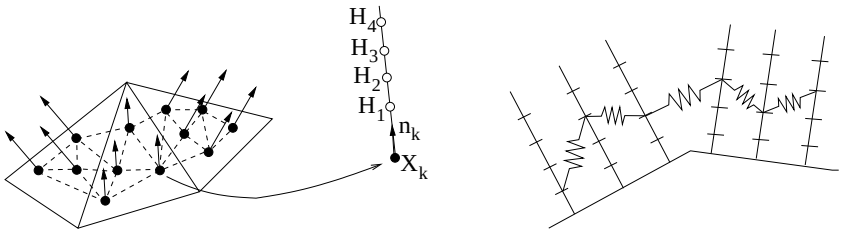

Figure 1: *The 3D* MRF *model. Left: Sample points* $\mathbf{X}_k$ *(black dots), are defined on a base surface and surface normals* $\mathbf{n}_k$*, are computed at those points. A neighbourhood relation* $\mathcal{N}$ *(dashed lines) is defined between the sample points. Labels* H_i *(white dots) are heights above the sample points. In the figure a set of 3 labels for a sample point are depicted, each of which corresponds to a 3D location in space. The cost of assigning a height to a sample point is based on the photo-consistency of the corresponding 3D location. Right: The smoothness cost involves terms proportional to distance between neighbouring relief surface points. The figure shows a 1D* MRF *where the smoothness cost forces minimum length. In the 2D case, an approximation to surface area is minimized.*

will lead to erroneous reconstructions if not all points are visible from all images. If we also require the base surface to be *outside* the true scene surface, as would be the case if it was obtained through the visual hull [3] for example, then it can be used as the occluding volume through which visibility can be inferred. In this case only positive heights (going *into* the volume) have to be examined. Such an occluding volume guarantees that no location in space *outside* or on the boundary of the volume is considered visible from an image if it is occluded by the true scene surface. On the other hand there may be visible locations that are erroneously considered occluded. For a proof of this claim see [13].

Note that the volume of the base surface cannot provide accurate information for the visibility of locations inside it. It can be used however as an approximation by assuming that $\mathbf{X}_k + h_k\mathbf{n}_k$ has the same visibility as $\mathbf{X}_k$ for the small range of heights we are considering. The base surface is therefore used to define a visibility map $V_n(\mathbf{X}_k)$ that is 1 when $\mathbf{X}_k$ is visible from image n and 0 otherwise. Taking this into account the labelling cost is set to

$$C_k(h_k) = w_1\rho\left\{\mathtt{i}_n(h_k) : \quad V_n(\mathbf{X}_k) = 1\right\}. \tag{3}$$

2.2 Compatibility cost

As mentioned previously, the dense stereo problem is ill posed and some form of regularization is necessary. In a 3D, non regular MRF, defining the notion of 'compatible' neighbouring heights presents a challenge. In the simple case where base surface normals are parallel (planar regions) and distances between sample points are constant, simple choices for the compatibility cost such as $\|h_k - h_l\|$ or $\|h_k - h_l\|^2$ work adequately. These costs also permit a significant speed up to the BP algorithm described in [6]. They are not very meaningful however for curved base surfaces where the distance between sample points and direction of surface normals need to be taken into account. The cost function

$$C_{kl}(h_k, h_l) = w_2 d_{kl}(h_k, h_l) \tag{4}$$

with some weight parameter w_2 and $d_{kl}(h_k, h_l) = \|(\mathbf{X}_k + h_k\mathbf{n}_k) - (\mathbf{X}_l + h_l\mathbf{n}_l)\|$, penalizes the Euclidean distance between neighbouring relief surface points. It favours minimal area surfaces and is meaningful for arbitrary configurations of base surface and sample points (fig. 1 right).

3 Optimisation

The MRF model laid out in the previous section provides a probability for any possible height labelling and corresponding relief surface. MRF inference involves recovering the most probable site labelling which is an NP-hard optimization problem in its generality [12]. Fortunately a number of efficient approximate algorithms have been proposed such as graph cuts [1] and belief propagation [20]. These methods have been shown to give very good results in a depth-map setting (see [18, 21] for a comparison). In this work we choose to apply a belief propagation scheme which we outline in the following section.

3.1 Loopy Belief Propagation

Belief propagation works by the circulation of messages across neighbouring sites. Each site sends to each of its neighbours a message with its belief about the probabilities of a neighbour being assigned a particular height. The clique potentials

$$\Phi_k(h_k) = \exp(-C_k(h_k)) \tag{5}$$

and

$$\Psi_{kl}(h_k, h_l) = \exp(-C_{kl}(h_k, h_l)) \tag{6}$$

are precomputed and stored as $L \times 1$ and $L \times N$ matrices respectively. Now suppose that $m_{ij}(h_j)$ denotes the message sent from sample point i to sample point j (this is a vector indexed by possible heights at j). We chose to implement the max-product rule according to which, after all messages have been exchanged, the new message sent from k to l is

$$\widetilde{m}_{kl} = \max_{h_k} \Phi_k(h_k)\Psi_{kl}(h_k, h_l) \prod_{i \in \mathcal{N}(k) - \{l\}} m_{ik}(h_k). \tag{7}$$

The update of messages can either be done synchronously after all messages have been transmitted, or asynchronously with each sample point sending messages using all the latest messages it has received. We experimented with both methods and found the latter to give speedier convergence, which was also reported in [21].

3.2 Coarse to fine strategy

One of the limitations of loopy belief propagation is that it has significant memory requirements, especially as the size of the set of possible heights is increased. In the near future bigger and cheaper computer memory will make this problem irrelevant, but for the system described in this paper we designed a simple coarse to fine strategy that allows for effective height resolutions of thousands of possible heights. This strategy effectively, instead of considering one BP problem with L different labels, considers $\log L / \log l$ problems with l labels where $l << L$. It therefore also offers a runtime speedup since it reduces the time required from $O(ML^2)$ to $O(\log L M l^2 / \log l)$.

Initially the label set for all sites corresponds to a coarse quantization of the allowable height range. After convergence of the Belief Propagation algorithm each site is assigned a label. In the next iteration a finer quantization of the heights is used within a range centered at the optimal label of the previous iteration. The label set is now allowed to be different for each site. At each phase the number of possible heights per node is constant but the height resolution increases.

To make this idea more precise, at this point we replace height labels with *height range* labels. A sample point can now be labelled by a height range in which its true height should lie. The cost for assigning height interval $[H_i, H_{i+1}]$ to the kth site is now defined as:

$$\hat{C_k}([H_i, H_{i+1}]) = \min_{h \in [H_i, H_{i+1}]} C_k(h). \tag{8}$$

In practice this minimum is computed by densely sampling $C_k(h)$ over the maximum range $[H_{min}, H_{max}]$ so that the images are all sampled at a sub-pixel rate. This computation only has to be performed at the beginning of the algorithm. Similarly the smoothness cost for assigning height ranges $[H_i, H_{i+1}], [H_j, H_{j+1}]$ to two neighbouring sites k and l is:

$$\hat{C}_{kl}([H_i, H_{i+1}], [H_j, H_{j+1}]) = C_{kl}\left(\tfrac{H_i + H_{i+1}}{2}, \tfrac{H_j + H_{j+1}}{2}\right) \ . \tag{9}$$

When belief propagation converges, each point is assigned an interval in which its height is most likely to lie. This interval will then be subdivided into smaller subintervals which become the site's possible labels. The process repeats until we reach the desired height resolution.

4 Results

In this section, a quantitative analysis using an artificial scene with ground truth is provided. Results on a challenging low-relief scene of a roman sarcophagus, a building facade and a stone carving are also illustrated. The weight parameters w_1 and w_2 of equations 3 and 4 are empirically set. However, in cases where the distributions of ρ and d_{kl} are known (e.g., we are given ground truth data for a similar scene), the weights can be set by using the approximation of [7] where the clique potentials are fitted to the distributions of ρ and d_{kl}.

4.1 Artificial scene

The artificial scene was a unit sphere whose surface was normally deformed by a random displacement and texture mapped with a random pattern (see fig. 2). The object was rendered from 20 viewpoints around the sphere. Using the non-deformed sphere as the base surface on which 40000 sample points were defined, the relief surface MRF was optimized by the method described in this paper (fig. 2). Positive and negative heights were considered but the visibility reasoning was still approximately correct because of the small height range considered. The performance of the relief surface approach was measured against a two-view Loopy Belief Propagation algorithm similar to the one described in [20]. To that end 10 pairs of nearby views were input to the BP algorithm resulting in 10 disparity maps. These maps were compared against the depth-maps of the reconstructed

	2-view BP	Relief Surf.
MSE	1.466 pixels	0.499 pixels
% of correct disparities	75.9%	79.1%

Table 1: *Artificial Scene. Comparison with 2-view BP. Both metrics show the superior performance of the relief surface approach. Note that a disparity estimate for a pixel is assumed* correct *if it is within one pixel of the true disparity.*

sphere from identical viewpoints. Table 4.1 shows the mean square errors of the two algorithms against the known ground truth. It also shows the percentage of correctly labelled pixels. Both figures demonstrate the superior performance of the relief surface approach which allows for simultaneous use of all data and for a viewpoint independent smoothness cost.

4.2 Real Scenes

For the first experiment presented here, three 1600×1200 pixel images of a Roman sarcophagus were used. The image regions of interest that were actually used for the reconstruction were approximately 600×300 pixels. The base surface was initialized to a rectangular planar region by manually clicking on four correspondences. A regular grid of 160000 sample points was then defined on this rectangle. The initial height range was subdivided by a factor of four in each stage of the coarse-to-fine scheme. The resulting height fields of the first three iterations are shown in fig. 4 where high intensity denotes positive height from the surface towards the viewer. Figure 3 shows textured and untextured versions of the reconstructed surface. Videos of these reconstructions can be found in `http://mi.eng.cam.ac.uk/~gv215/relsurf`.

The second experiment (fig. 5) was performed on three images of a building facade which the shiny or transparent windows make particularly difficult. The base surface was again a hand-initialized plane. Finally the third experiment was performed on three images of a stone carving. To illustrate the effect of a more complex but still approximate base surface, a sparse set of feature matches was Delaunay triangulated to obtain a base surface as a mesh. The relief surface was then optimized to yield the results shown in fig. 6.

5 Conclusion

In this paper we have shown how MRF techniques for image based stereo can be extended in the volumetric stereo domain. This is done by defining a set of sample points on a coarse base surface, establishing an MRF on unknown displacements of these points normal to the base surface. By casting the problem in the MRF framework we can use computationally tractable algorithms like belief propagation to recover the unknown displacements. Additionally, this parameterization of the scene is more general than a depth map and leads to image and viewpoint independent reconstructions. The MRF's compatibility cost favours solutions with minimal surface area. Furthermore, the base surface can be used as the occluding volume through which the visibility of individual sample

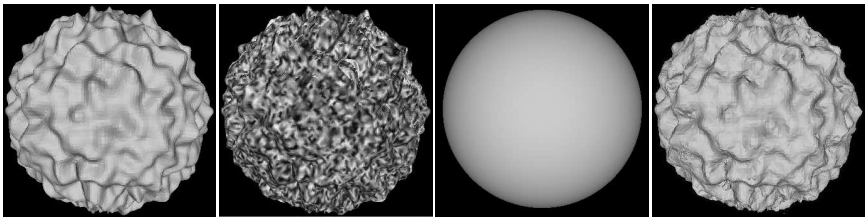

Figure 2: *Artificial Scene. From left to right: (a) The true scene (a unit sphere whose surface is deformed by a random positive or negative normal displacement). (b) The deformed sphere is texture mapped with a random pattern. (c) The base surface (a non deformed unit sphere). (d) The relief surface returned by the algorithm.*

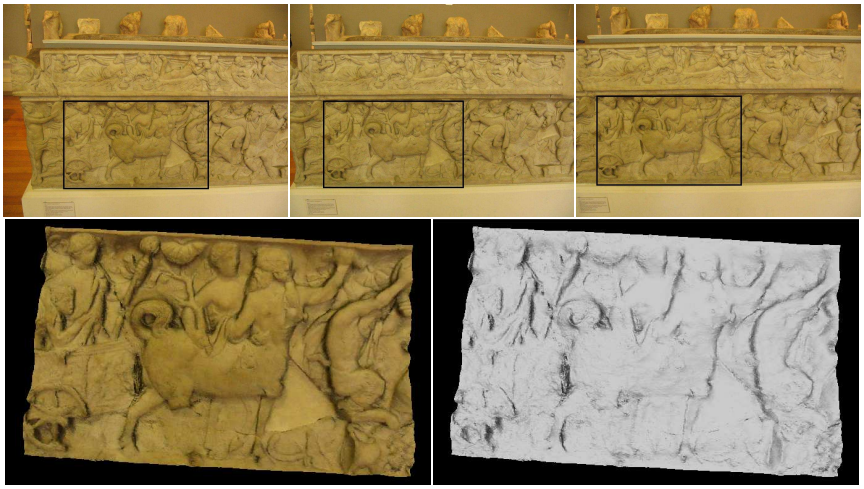

Figure 3: *Roman sarcophagus. Top: the three images used in the reconstruction with region of interest denoted by a black box. Bottom left: texture mapped rendering of reconstructed relief surface. Bottom right: without texture mapping. The base surface was a plane.*

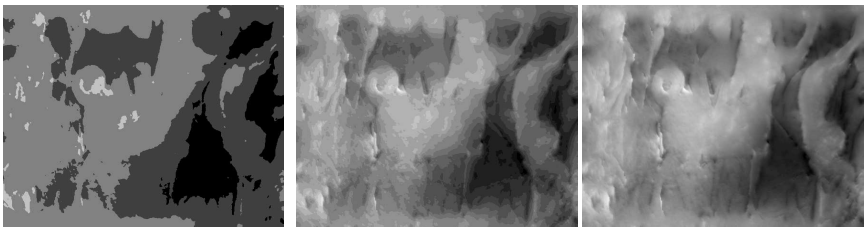

Figure 4: *Detail of the coarse to fine strategy. This is the output of the first three phases of the algorithm for the first experiment. The resolutions at each phase are 4, 16 and 64 height ranges shown from left to right.*

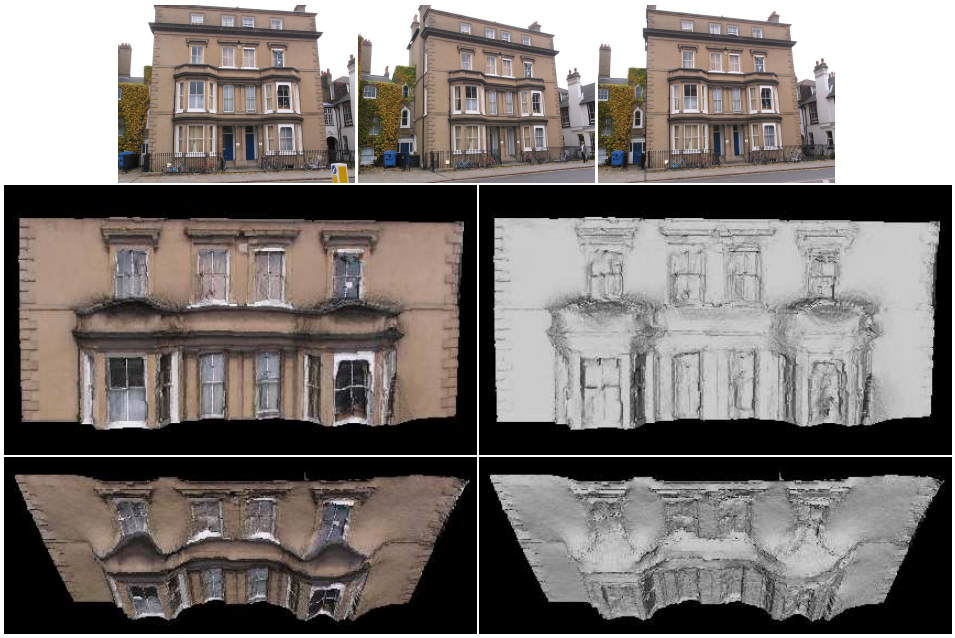

Figure 5: *Building facade. Top: the images used. Bottom two rows, left and right: texture mapped and untextured relief surface. The base surface was the wall plane. The challenge of the scene is the shiny or transparent windows as well as the fine relief at places.*

Figure 6: *Stone carving. Top: the images used. Bottom left: the base surface. Bottom middle: the untextured relief surface. Bottom right: the texture mapped relief surface.*

points is inferred. The memory requirements of belief propagation are reduced through the employment of a novel coarse-to-fine scheme. Promising results are demonstrated on a variety of real world scenes.

Acknowledgements

This work is supported by the Gates Cambridge Trust and Toyota Corporation.

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *Proceedings of CVPR 1998*, pages 648–655, 1998.

[2] R. Cipolla, Y. Okamoto, and Y. Kuno. Robust structure from motion using motion parallax. In *ICCV 93*, pages 374–382, 1993.

[3] G. Cross and A. Zisserman. Surface reconstruction from multiple views using apparent contours and surface texture. In *NATO Adv. Research Workshop on Confluence of C. Vision and C. Graphics, Ljubljana, Slovenia*, pages 25–47, 2000.

[4] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. *Computer Graphics*, 30(Annual Conference Series):11–20, 1996.

[5] O. Faugeras and R. Keriven. Variational principles, surface evolution, pdes, level set methods and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):335–344, 1998.

[6] P. F. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. In *To appear in Proceedings of CVPR 2004*, 2004.

[7] W. Freeman and E. Pasztor. Learning to estimate scenes from images. In M. Kearns, S. Solla, and D Cohn, editors, *Adv. Neural Information Processing Systems*, volume 11. MIT Press, 1999.

[8] P. Fua and Y. G Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *IJCV*, 16:35–56, September 1995.

[9] J. Hailin, S. Soatto, and A.J. Yezzi. Multi-view stereo beyond lambert. In *CVPR 2003*, volume 1, pages 171–178, june 2003.

[10] J. Isidoro and S. Sclaroff. Stochastic refinement of the visual hull to satisfy photometric and silhouette consistency constraints. In *Proc. Int. Conf. on Computer Vision*, pages 1335–1342, 2003.

[11] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proceedings of ICCV 2001*, pages 508–515, 2001.

[12] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph-cuts. In *ECCV 2002*, volume 3, pages 82–96, 2002.

[13] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.

[14] A. Lee, H. Moreton, and H. Hoppe. Displaced subdivision surfaces. In *Siggraph 2000, Computer Graphics Proceedings*, pages 85–94, 2000.

[15] P.J. Narayanan, P.W. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *ICCV98*, pages 3–10, 1998.

[16] S. Paris, F. Sillion, and L. Quan. A surface reconstruction method using global graph cut optimization. In *Proceedings of Asian Conference on Computer Vision*, January 2004.

[17] S. Roy and I. J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *Proceedings of ICCV 1998*, pages 735–743, 1998.

[18] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1–3):7–42, 2002.

[19] J. W. Shade, S. J. Gortler, L. -W. He, and R. Szeliski. Layered depth images. *Computer Graphics*, 32(Annual Conference Series):231–242, 1998.

[20] J. Sun, H,-Y Shum, and N. -N. Zheng. Stereo matching using belief propagation. In *Proceedings of ECCV,2002*, pages 510–524, 2002.

[21] F. M. Tappen and W. T. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *ICCV 2003*, volume 2, pages 900–907, 2003.

[22] A. Treuille, A. Hertzmann, and S. Seitz. Example-based stereo with general brdfs. In *8th European Conference on Computer Vision (ECCV 2004)*, may 2004.

[23] L. Zhang and S. M. Seitz. Image-based multiresolution shape recovery by surface deformation. In *Proc. SPIE: Videometrics and Optical Methods for 3D Shape Measurement*, pages 51–61, 2001.