Multiview Stereo via Volumetric Graph-Cuts and Occlusion Robust Photo-Consistency

George Vogiatzis, *Member*, *IEEE*,
Carlos Hernández, *Member*, *IEEE*,
Philip H.S. Torr, *Member*, *IEEE*, and
Roberto Cipolla, *Member*, *IEEE*

Abstract—This paper presents a volumetric formulation for the multiview stereo problem which is amenable to a computationally tractable global optimization using Graph-cuts. Our approach is to seek the optimal partitioning of 3D space into two regions labeled as "object" and "empty" under a cost functional consisting of the following two terms: 1) A term that forces the boundary between the two regions to pass through photo-consistent locations and 2) a ballooning term that inflates the "object" region. To take account of the effect of occlusion on the first term, we use an occlusion robust photo-consistency metric based on Normalized Cross Correlation, which does not assume any geometric knowledge about the reconstructed object. The globally optimal 3D partitioning can be obtained as the minimum cut solution of a weighted graph.

Index Terms—3D/stereo scene analysis, shape, graph algorithms, global optimization.

1 Introduction

This paper considers the problem of reconstructing the dense geometry of a 3D object from a number of images in which the camera pose and intrinsic parameters have been previously obtained. This is a classic computer vision problem that has been extensively studied and a number of solutions have been published. Work in the field can be categorized according to the geometrical representation of the 3D object with the majority of papers falling under one of the following two categories: 1) algorithms that recover depth-maps with respect to an image plane and 2) volumetric methods that represent the volume directly, without any reference to an image plane.

In the first class of methods, 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. An excellent review for image-based methods can be found in Scharstein and Szeliski [22]. 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 using Graph-cuts [5], [14], [21] or Loopy Belief Propagation [27]. They can also be formulated as continuous PDE evolutions on the depth maps [26]. 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 function of image point. Capturing complete objects in this manner requires further processing to merge multiple depth maps. This was recently attempted in [10] but resulted in only partially reconstructed object surfaces, leaving holes in areas of uncertainty. A second limitation is that the smoothness term imposed by the MRF is defined on image disparities or depths and, hence, is viewpoint dependent, i.e., if a different view is chosen as the reference image the results may be different.

The second class comprises of methods that use a *volumetric representation* of shape. For a recent, very thorough review of related techniques, see [23]. Under this framework, multiple viewpoints can be easily integrated and surface smoothness can be enforced independent of viewpoint. This class consists of techniques using implicit representations such as voxel occupancy grids [16], or level-sets of 3D scalar fields [7], [20], and explicit representations such as polygonal meshes [8], [11]. While some of these methods are known to produce high quality reconstructions, their convergence properties in the presence of noise are not well understood. Due to lack of regularization, methods based on Space Carving [16] produce surfaces that tend to *bulge out* in regions of low surface texture (see the discussion about shape priors in [23]). In variational schemes such as level-sets and mesh-based stereo, the optimal surface is usually obtained via gradient descent optimization. As a result, these techniques typically employ multiresolution coarse-to-fine strategies to decrease the probability of getting trapped in local minima (e.g., [7], [8], [11], [20]). Furthermore, explicit representations such as meshes are known to suffer from topological and sampling problems [18].

The approach described in this paper combines the advantages of both classes described above. We adopt an implicit volumetric representation based on voxel occupancy, but we pose the reconstruction problem as finding the minimum cut of a weighted graph. This computation is exact and can be performed in polynomial time. The benefits of our approach are the following:

1. Objects of arbitrary topology can be fully represented and computed as a single surface with no self-intersections.
2. The representation and geometric regularization is image and viewpoint independent.
3. Global optimization is computationally tractable, using existing max-flow algorithms.

1.1 Background and Previous Work

The inspiration for the approach presented in this paper is the work of Boykov and Kolmogorov [2], which establishes a theoretical link between maximum flow problems in discrete graphs and minimal surfaces in an arbitrary Riemannian metric. In particular, the authors show how a continuous Riemannian metric can be approximated by a discrete weighted graph so that the max-flow/min-cut solution for the graph corresponds to a local geodesic or minimal surface in the continuous case. The application described in that paper is interactive 2D or 3D segmentation. A probabilistic formulation of interactive segmentation with a more elaborate foreground/background model was given in Blake et al [1].

In [29], we showed how the basic idea of [2] can be applied to the volumetric, multiview stereo problem by computing a photo-consistency-based Riemannian metric in which a minimal surface is computed. In that method, two basic assumptions are made: First, it is assumed that the object surface lies between two parallel boundary surfaces. The outer boundary is usually obtained from the visual hull while the inner boundary lies at a constant distance inside the outer boundary. This effectively limits the depth of concavities that can be represented in the reconstructed object. The second assumption is that the visibility of each point on the object's surface can be determined from the visibility of the closest point on the outer surface. Even though both of these assumptions are satisfied for a large class of objects and acquisition set-ups, they restrict the applicability of the method considerably. Nevertheless, by demonstrating promising results and highlighting the feasibility of solving multiview stereo using volumetric graph cuts, [29] inspired a number of techniques [4], [9], [13], [17], [24], [25], [28] that built on our formulation and attempted to address some of its shortcomings.

In Furukawa and Ponce [9] and Sinha and Pollefeys [24], two different ways were proposed for incorporating the powerful silhouette cue into the graph-cut framework while Starck et al.

- *G. Vogiatzis and C. Hernández are with Toshiba Research Europe Ltd., 208 Cambridge Science Park, Milton Road, Cambridge CB4 0GZ, UK. E-mail: {george.vogiatzis, carlos.hernandez}@crl.toshiba.co.uk.*
- *P.H.S. Torr is with Brookes Computer Vision Group, Department of Computing, Oxford Brookes University, Wheatley, Oxford OX33 1HX, UK. E-mail: philiptorr@brookes.ac.uk.*
- *R. Cipolla is with the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK. E-mail: cipolla@eng.cam.ac.uk.*

Manuscript received 26 Oct. 2006; revised 23 Mar. 2007; accepted 29 Mar. 2007; published online 6 June 2007.
Recommended for acceptance by R. Zabih.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0757-1006.
Digital Object Identifier no. 10.1109/TPAMI.2007.70712.

0162-8828/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

TABLE 1
Comparison of Our Method with State-of-the-Art Techniques against Ground Truth Data (from [23])

	Accuracy / Completeness		
	Full (312 images)	Ring (47 images)	SparseRing (16 images)
Hernandez [11]	0.36mm / 99.7%	0.52mm / 99.5%	0.75mm / 95.3%
Goesele [10]	0.42mm / 98.0%	0.61mm / 86.2%	0.87mm / 56.6%
Hornung [13]	0.58mm / 98.7%	–	–
Pons [20]	–	0.60mm / 99.5%	0.90mm / 95.4%
Furukawa [9]	0.65mm / 98.7%	0.58mm / 98.5%	0.82mm / 94.3%
Vogiatzis [29]	1.07mm / 90.7%	0.76mm / 96.2%	2.77mm / 79.4%
Present method	**0.50mm / 98.4%**	**0.64mm / 99.2%**	**0.69mm / 96.9%**

[25] and Tran et al. [28] showed how to enforce sparse feature matches as hard constraints. Hornung and Kobbelt [13] improved the construction of the voxel grid and cast the method in a hierarchical framework that allows for a significant speedup at the expense of no longer obtaining a global optimum. Finally, Lempitsky et al. [17] offer an alternative approach for visibility reasoning, while in [4] this is expanded to incorporate the idea of *photo-flux* as a data-driven ballooning force that helps reconstruct thin protrusions and concavities. Additionally, [4] and [17] were the first papers to propose a global optimization scheme for volumetric multiview stereo that did not require any initialization (e.g., visual hull). However, the reconstructions shown were less detailed than those obtained with other state-of-the-art techniques and no comparison or quantitative analysis was provided.

In this paper, we improve the original formulation of the method of [29] by relaxing the two assumptions described above. Hence, in the present formulation, 1) the object surface is not geometrically constrained to lie between an inner and an outer surface and 2) no explicit reasoning about visibility is required. This is achieved through the use of a robust shape-independent photo-consistency cost first used in [11]. The key idea behind that scheme is that occluded pixels are treated as outliers in the matching process. Furthermore, the formulation presented here achieves reconstruction results of far superior accuracy than [29], as demonstrated by results from a scene where ground truth is available (Fig. 5 and Table 1).

The rest of the paper is laid out as follows: Section 2 describes how multiview stereo can be formulated as a graph-cut optimization. In Section 3, we describe the photo-consistency functional associated with any candidate surface while Section 4 explains how this functional is approximated with a discrete flow graph. Section 5 presents our 3D reconstruction results on real objects and Section 6 concludes with a discussion of the paper's main contributions.

2 Graph-Cuts for Volumetric Stereo

In [2], and, subsequently, in [1], it was shown how graph-cuts can optimally partition 2D or 3D space into "foreground" and "background" regions under any cost functional consisting of the following two terms:

- **Foreground/background cost.** For every point in space, there is a cost for it being "foreground" or "background."
- **Discontinuity cost.** For every point in space, there is a cost for it lying on the boundary between the two partitions.

Mathematically, the cost functional described above can be seen as the sum of a weighted *surface area* of the boundary surface and a weighted *volume* of the "foreground" region as follows:

$$E[S] = \iint_S \rho(\mathbf{x}) dA + \iiint_{V(S)} \sigma(\mathbf{x}) dV, \qquad (1)$$

where S is the boundary between "foreground" and "background," $V(S)$ denotes the "foreground" volume enclosed by S and ρ, and σ are two scalar density fields.

The application described in [2] was the problem of 2D/3D segmentation. In that domain, $\rho(\mathbf{x})$ is defined as a function of the image intensity gradient and $\sigma(\mathbf{x})$ as a function of the image intensity itself or local image statistics. In this paper, we show how multiview stereo can also be described under the same framework with the "foreground" and "background" partitions of 3D space corresponding to the reconstructed object and the surrounding empty space, respectively.

Our model balances two competing terms: The first one minimizes a surface integral of photo-consistency while the second one maximizes volume. The following two sections describe the two terms of our multiview stereo cost functional in more detail.

2.1 Foreground/Background Cost

A challenge specific to the multiview stereo problem, is that there is no straightforward way to define the foreground/background model $\sigma(\mathbf{x})$. This is because in this problem our primary source of geometric information is the *correspondence cue* which is based on the following observation: A 3D point located *on* the object surface projects to image regions of *similar* appearance in all images where it is not occluded. Using this cue, one can label 3D points as being *on* or *off* the object surface but cannot directly distinguish between points *inside* or *outside* it. In contrast, the *silhouette cue* is based on the requirement that all points *inside* the object volume must project inside the silhouettes of the object that can be extracted from the images. Hence, the silhouette cue can provide some foreground/background information by giving a very high likelihood of being *outside* the object to 3D points that project outside the silhouettes. In [4] a data driven, foreground/background model based on the concept of *photo-flux* has been introduced. To compute photo-flux, surface orientation must be either estimated (in the case of global optimization) or the *current* surface orientation is used (in the case of gradient-descent surface evolution).

In this work, we adopt a very simple, data-independent model where $\sigma(\mathbf{x})$ is defined as a negative constant λ that produces an inflationary (*ballooning*) tendency. The motivation for this type of term in the active contour domain is given in [6], but intuitively, it can be thought of as a shape prior that favors objects that fill the bounding volume in the absence of any other information. If the value of λ is too large, then the solution tends to overinflate, filling the entire bounding volume while if λ is too small then the solution collapses into an empty surface. For values of λ in between these two cases, the algorithm converges to the desired surface. In practice, it is quite easy to find a value of λ which will work by performing a few trial runs. As there is a large range of suitable λ values, all of which give nearly identical results, no detailed search for the optimal λ value is necessary.

Additionally, we can encode any silhouette information that may be available by setting $\sigma(\mathbf{x})$ to be infinitely large when $\mathbf{x}$ is outside the visual hull. Furthermore, if we can also assume, as in [29], that the concavities of the object are of a maximum depth D from the visual hull then we can set $\sigma(\mathbf{x})$ to be infinitely small when $\mathbf{x}$ is inside the visual hull at a distance, at least D from it. In many cases, such as the experiments of Figs. 1 and 4 where the objects have relatively simple topology, a bounding box guaranteed to contain the object is sufficient to obtain a good reconstruction. To encode this knowledge, we just need to set $\sigma(\mathbf{x})$ to be infinitely large when $\mathbf{x}$ is outside that bounding box.

2.2 Discontinuity Cost

The second challenge of multiview stereo is that the surface area density ρ, which corresponds to the discontinuity cost, is a function of the photo-consistency of the point in space, which, in turn, depends on which cameras are visible from that point. Consequently in multiview stereo, the discontinuity cost has the form $\rho(\mathbf{x}, S)$ since the surface S itself determines camera visibility. The graph-cut formulation of [2] cannot easily be adapted to cope with

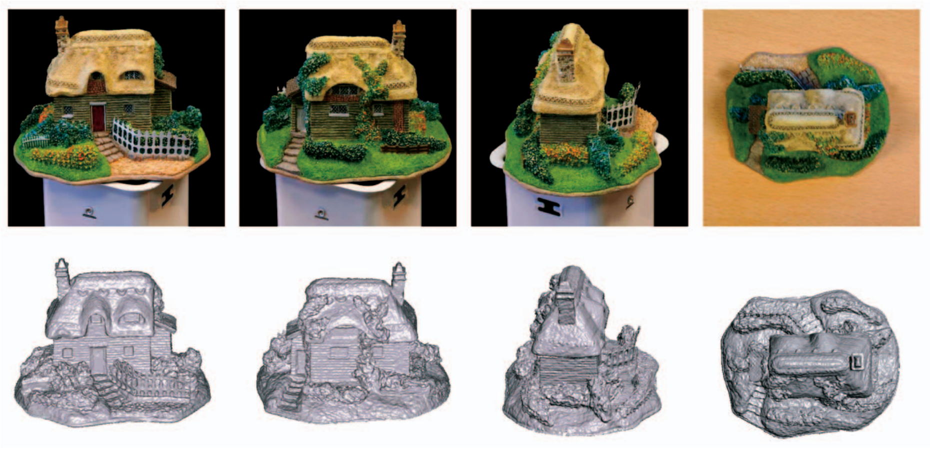

Fig. 1. **Toy House.** This is an example of a 3D model of a real object, obtained using the technique described in this paper. In the top row are four images of a toy house while in the bottom row, renderings of the 3D model from similar viewpoints are shown. The first three images were part of the input sequence used while the fourth was not shown to the algorithm. The model of this small toy house (approximately 10 cm in diameter) contains accurately reconstructed submillimeter details such as the fence and the relief of the roof.

this type of cost functional. In [13], [29], the problem is solved by assuming the existence of an approximate surface S_{approx}, provided by the visual hull or otherwise, which provides visibility information. However, as self-occlusions not captured by the approximate surface will be ignored, the accuracy of the results may suffer. Also, such approximate object surface may not be readily available. Our approach is to use a photo-consistency metric that accounts for occlusions using robust Normalized Cross-Correlation (NCC) voting without any dependence on approximate object geometry. The surface cost functional that we optimize is

$$E[S] = \iint_S \rho(\mathbf{x}) dA - \lambda \iiint_{V(S)} dV. \tag{2}$$

The next section will describe the photo-consistency metric $\rho(\mathbf{x})$ in more detail.

3 Photo-Consistency Metric

The input to our method is a sequence of images $I_1, \dots, I_N$ calibrated for camera pose and intrinsic parameters. The photo-consistency of a potential scene point $\mathbf{x}$ can be evaluated by comparing its projections in the images where it is visible. We propose the use of a robust photo-consistency metric similar to the one described in [11] that does not need any visibility computation. This choice is motivated by the excellent results obtained by this type of photo-consistency metric in the recent comparison of 3D modeling techniques carried out by [23]. The basic idea is that all potential causes of mismatches like occlusion, image noise, lack of texture, or highlights are uniformly treated as outliers in the matching process. Matching is then seen as a process of robust model fitting to data containing outliers. Specifically, for a given 3D point $\mathbf{x}$, its photo-consistency value $\rho(\mathbf{x})$ is computed by asking every image i to give a vote for that location. Specifically, we define

$$\rho(\mathbf{x}) = \exp\left\{ -\mu \sum_{i=1}^{N} \mathrm{VOTE}_i(\mathbf{x}) \right\}, \tag{3}$$

where μ is very stable rate-of-decay parameter which in all our experiments was set to 0.05.

The value of $\mathrm{VOTE}_i(\mathbf{x})$ is computed as follows:

- Compute the corresponding optic ray

  $$\mathbf{o}_i(d) = \mathbf{x} + (\mathbf{c}_i - \mathbf{x})d \tag{4}$$

  that goes through the camera's optic center $\mathbf{c}_i$ and the 3D point $\mathbf{x}$.
- As a function of the depth along the optic ray d, project the 3D point $\mathbf{o}_i(d)$ into the M closest cameras $\mathcal{N}(i)$ and compute M correlation scores $S_j(d)$ between image $I_{j\in\mathcal{N}(i)}$ and the reference image I_i. Each score $S_j(d)$ is obtained using normalized cross correlation between two square windows centered on the projections of $\mathbf{o}_i(d)$ into I_i and $I_{j\in\mathcal{N}(i)}$. For the experiments presented here we used 11×11 pixel windows.
- combine the M correlation scores $S_j(d)$ into a single score $\mathcal{C}(d)$, and give a vote to the 3D location $\mathbf{x}$, i.e., $\mathbf{o}_i(0)$, only if $\mathcal{C}(0)$ is the global maximum of $\mathcal{C}$ as follows:

$$\mathrm{VOTE}_i = \begin{cases} \mathcal{C}(0) & if \quad \mathcal{C}(0) \geq \mathcal{C}(d) \ \forall d \\ 0 & otherwise. \end{cases} \tag{5}$$

One of the simplest ways of combining the M correlation scores for every depth d is to simply average them, i.e.,

$$\mathcal{C}(d) = \sum_{j\in\mathcal{N}(i)} S_j(d). \tag{6}$$

However, averaging does not allow the robust handling of occlusions, highlights, or lack of texture. In order to obtain a better score $\mathcal{C}(d)$, we make an important observation: Because of different types of noise in the image, the global maximum of a single correlation curve does not always correspond to the correct depth. However, if the surface is seen by the camera without occlusion or sensor saturation, the correlation score **does** show a local maximum near the correct depth, though it may not be the global one. In order to take into account this observation, we build a new $\mathcal{C}$ by detecting all the local maxima d_k of S_j, i.e., $\frac{\partial S_j}{\partial d}(d_k) = 0$, $\frac{\partial^2 S_j}{\partial d^2}(d_k) > 0$, and using a Parzen window [19] with a kernel W as follows:

$$\mathcal{C}^w(d) = \sum_{j\in\mathcal{N}(i)} \sum_k S_j(d_k) W(d - d_k). \tag{7}$$

The Parzen window technique provides an effective way of taking into account the actual scores of the local maxima **and** reinforcing those local maxima that are close to each other. It provides very good robustness against occlusion and image noise, which in practice makes it the core of a photo-consistency measure that does not need explicit visibility computation. Fig. 2 demonstrates the benefits of the Parzen filtering technique as opposed to simple averaging of correlation scores. For the example of Fig. 2, a Gaussian kernel has been used. In practice, we discretize the 3D volume into voxels and we count the number of local maxima that fall inside a voxel. This corresponds to using a rectangular kernel with width equal to the size of the voxel grid.

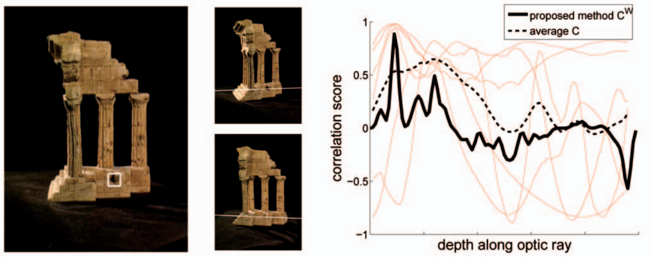

Fig. 2. **Robust voting versus averaging.** Our algorithm robustly estimates the depth of a pixel in an input image (left) by computing NCC scores between a patch centered on that pixel and patches along points on corresponding epipolar lines in the M closest images, two of which are shown in the middle column. In this way, M correlation curves are obtained (in our example $M = 6$). These curves are plotted here in red across depth along the optic ray. Curves corresponding to unoccluded viewpoints (such as the top-middle image) share a local optimum in the same location which corresponds to the correct surface depth. Curves from occluded viewpoints (such as the bottom-middle image) do not have an optimum in that location and, hence, a simple averaging of the curves (dashed line) does not work. By computing a sliding Parzen filter on the local maxima of the correlation curves (here, we have used a Gaussian kernel), the correct depth can be recovered at the point of maximum response.

4 Graph Structure

To obtain a discrete solution to (2), 3D space is quantized into voxels of size $h \times h \times h$. The graph nodes consist of all voxels whose centers are within a certain bounding box that is guaranteed to contain the object. For the results presented in this paper, these nodes were connected with a regular 6-neighborhood grid. Bigger neighborhood systems can be used which provide a better approximation to the continuous functional (2), at the expense of using more memory to store the graph. Now, assume two voxels centered at $\mathbf{x_i}$ and $\mathbf{x_j}$ are neighbors. Then, the weight of the edge joining the two corresponding nodes on the graph will be [2]

$$w_{ij} = \frac{4\pi h^2}{3}\rho\Big(\frac{\mathbf{x_i}+\mathbf{x_j}}{2}\Big), \tag{8}$$

where $\rho(\mathbf{x})$ is the matching cost function defined in (3). In addition to these weights between neighboring voxels, there is also the ballooning force edge connecting every voxel to the source node with a constant weight of $w_b = \lambda h^3$. Finally, the outer voxels that are part of the bounding box (or the voxels outside the visual hull if that is available) are connected with the sink with edges of infinite weight. The configuration of the graph is shown in Fig. 3b.

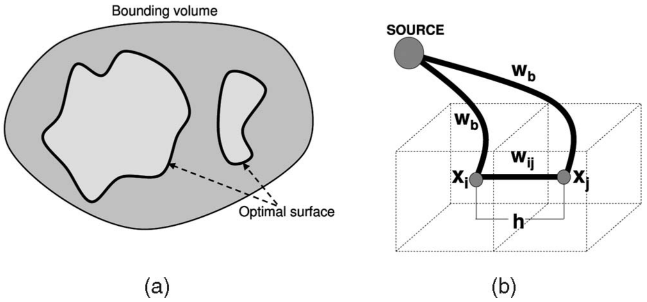

Fig. 3. **Surface geometry and flow graph construction.** (a) A 2D slice of space showing the bounding volume and the optimal surface inside it that is obtained by computing the minimum cut of a weighted graph. Note that complicated topologies such as holes or disjoint volumes can be represented by our model and recovered after optimization. (b) The correspondence of voxels with nodes in the graph. Each voxel is connected to its neighbors as well as to the source.

It is worth pointing out that the graph structure described above can be thought of as a simple binary MRF. Variables correspond to voxels and can be labeled as being *inside* or *outside* the scene. The unitary clique potential is just 0 if the voxel is outside and w_b if it is inside the scene while the pairwise potential between two neighbor voxels i and j is equal to w_{ij} if the voxels have opposite labels and 0 otherwise. As a binary MRF with a *submodular* energy function [15] it can be solved exactly in polynomial time using Graph-cuts.

5 Results

In this section, we present some 3D reconstruction results obtained by our technique. The system used for all the models shown was a Linux-based Intel Pentium IV with 2GB RAM and running at 3.0 GHz. The spatial resolution for the voxel grids was 300^3 voxels for the toy house sequence (Fig. 1), 200^3 voxels for the Hygeia sequence (Fig. 4) and 256^3 v voxels for the Temple sequence (Fig. 5). The ballooning parameter λ was set to values between 0.1 and 1.0. Computation time is strongly dominated by the photo-consistency cost calculation which takes between 30 minutes and 1.5 hours depending on number of images and their resolution. Generally, the computational complexity of this part of the algorithm grows linearly with the total number of pixels in the sequence. The computation time required by the graph-cut computation for a 300^3 grid is approximately 45 minutes. We used the graph-cut

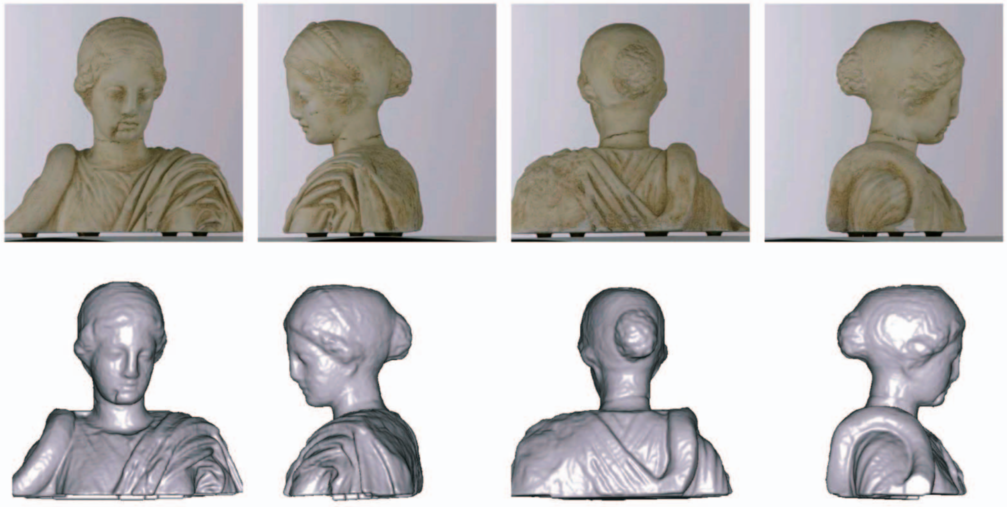

Fig. 4. **Reconstruction results.** Reconstruction of plaster bust of Greek goddess Hygeia. The input sequence consists of 36 images. Four of these are shown in the first row while the second row shows similar views of the reconstructed model.

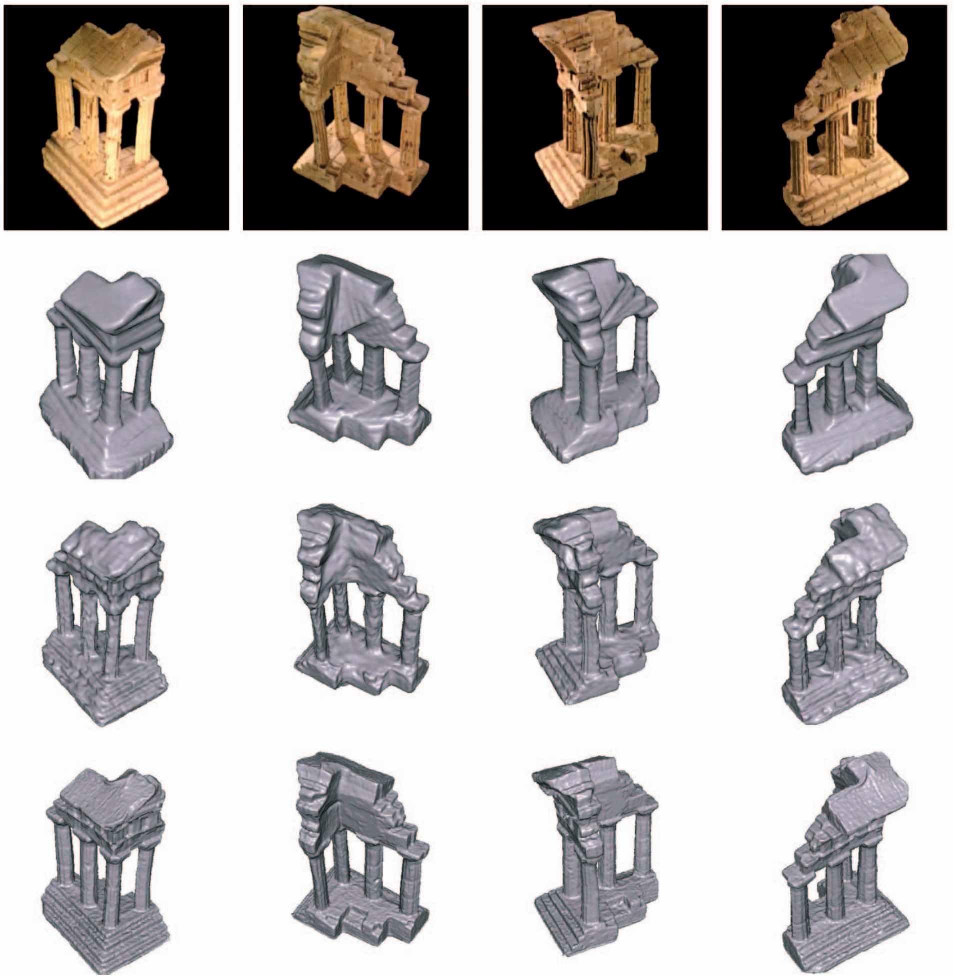

Fig. 5. **Castor and Pollux (Dioscuri) temple sequence.** First row: Four of the input images. Second row: Visual hull obtained from silhouettes. Third row: Results obtained with the original Volumetric Graph-cuts formulation of [29]. Fourth row: Results obtained with the method presented here. The occlusion robust photo-consistency metric greatly enhances the detail of the reconstruction.

algorithm proposed in [3] and in particular the implementation available at the authors' Web site.

The first experiment was performed on a plaster bust of the Greek goddess Hygeia (36 images) photographed with a 5M pixel digital camera. The object was mounted on a turntable and camera pose was obtained automatically using the object's silhouettes [12]. Note, however, that these silhouettes were not used for any other computation such as visual hull construction. The reconstruction results are shown in Fig. 4.

Our second experiment (Fig. 5) used images of a replica of the Castor and Pollux (Dioscuri) temple in Agrigento, Sicily, with a resolution of 640×480 pixels. Four of these images are shown in the first row of Fig. 5. This sequence was used as part of a multiview stereo evaluation effort which was presented in [23]. Camera motion is known and ground truth is available through the use of a laser scanner device (see [23] for details). Three different subsets of the sequence each with a different number of images are provided: The full set of 312 images (Full), a medium-sized sequence with 47 images (Ring), and a sparse sequence with only 16 images (SparseRing). As the object is photographed against a black background, silhouettes can be computed by simple thresholding. The visual hull obtained from those silhouettes is shown in the second row of Fig. 5. We have encoded this in our foreground/background term as described in Section 2.1. Fig. 5 shows the results of our reconstruction for the Full subsequence (fourth row) compared to the results obtained using the original formulation of Volumetric Graph-cuts [29] (third row). The improvement in geometric accuracy is especially evident in the rear view of the temple where, due to self-occlusions the visibility assumptions of [29] were severely violated. Our present formulation makes no such visibility approximations and, hence, is able to fully extract the geometry information contained in the images.

Fig. 6 provides a qualitative demonstration of the difference in discriminative power between the photo-consistency metric of [29] (Fig. 6a) and our current method (Fig. 6c). The figure shows slices of the two photo-consistency fields corresponding to the upper part of the temple above the columns. It demonstrates a significant reduction in photo-consistency noise brought about by the robust voting scheme of Section 3.

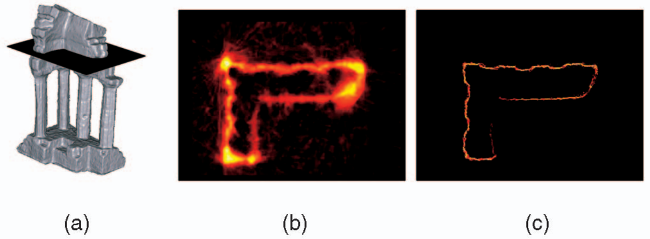

Fig. 6. **Noise reduction in photo-consistency.** (a) A slice of the photo-consistency volume taken through the entablature of the temple. (b) The metric of [29] contains falsely photo-consistent regions (e.g., near the corners). (c) The occlusion robust metric proposed here significantly suppresses noise and the correct surface can be accurately localized.

A quantitative analysis of our results and comparison with state-of-the-art techniques across all three subsequences is presented in Table 1. The accuracy metric shown is the distance d (in millimeters) that brings 90 percent of the reconstructed surface within d from some point on the ground truth surface. The completeness figure measures the percentage of points in the ground truth model that are within 1.25 mm of the reconstructed model. Under both metrics our method currently ranks among the top performers. In the SparseRing sequence with only 16 images, our method performs best in terms of both accuracy and completeness.

The final example, shown in Fig. 1 is from a high-resolution sequence of 140 images ($3,456 \times 2,304$ pixels) of a toy house of about 10 cm diameter. Camera calibration has been obtained automatically using silhouettes [12]. As in the first experiment, however, we did not include these silhouettes in our foreground/background term. The mesh obtained from the 300^3 voxel grid contains accurately reconstructed submillimeter details.

6 Discussion

This paper introduces the use of graph-cut optimization to the volumetric multiview stereo problem. We begin by defining an occlusion-robust photo-consistency metric which is then approximated by a discrete flow graph. This metric uses a robust voting scheme that treats pixels from occluded cameras as outliers. We then show how graph-cut optimization can exactly compute the *minimal* surface that encloses the largest possible volume, where *surface area* is just a surface integral in this photo-consistency field. The experimental results presented, demonstrate the benefits of combining a volumetric surface representation with a powerful discrete optimization algorithm such as graph-cuts.

References

[1] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive Image Segmentation Using an Adaptive GMMRF Model," *Proc. European Conf. Computer Vision,* pp. 428-441, 2004.

[2] Y. Boykov and V. Kolmogorov, "Computing Geodesics and Minimal Surfaces via Graph Cuts," *Proc. Int'l Conf. Computer Vision,* pp. 26-33, 2003.

[3] Y. Boykov and V. Kolmogorov, "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 9, pp. 1124-1137, Sept. 2004.

[4] Y. Boykov and V. Lempitsky, "From Photohulls to Photoflux Optimization," *Proc. British Machine Vision Conf.,* pp. 1149-1158, 2006.

[5] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 11, pp. 1222-1239, Nov. 2001.

[6] L.D. Cohen and I. Cohen, "Finite-Element Methods for Active Contour Models and Balloons for 2-D and 3-D Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 11, pp. 1131-1147, Nov. 1993.

[7] O. Faugeras and R. Keriven, "Variational Principles, Surface Evolution, PDES, Level Set Methods and the Stereo Problem," *IEEE Trans. Image Processing,* vol. 7, no. 3, pp. 335-344, 1998.

[8] P. Fua and Y.G. Leclerc, "Object-Centred Surface Reconstruction: Combining Multi-Image Stereo and Shading," *Int'l J. Computer Vision,* vol. 16, no. 1, pp. 35-56, 1995.

[9] Y. Furukawa and J. Ponce, "Carved Visual Hulls for Image-Based Modeling," *Proc. European Conf. Computer Vision,* vol. 1, pp. 564-577, 2006.

[10] M. Goesele, B. Curless, and S.M. Seitz, "Multi-View Stereo Revisited," *Proc. Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 2402-2409, 2006.

[11] C. Hernández and F. Schmitt, "Silhouette and Stereo Fusion for 3D Object Modeling," *Computer Vision and Image Understanding,* vol. 96, no. 3, pp. 367-392, 2004.

[12] C. Hernández, F. Schmitt, and R. Cipolla, "Silhouette Coherence for Camera Calibration under Circular Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 2, pp. 343-349, Feb. 2007.

[13] A. Hornung and L. Kobbelt, "Hierarchical Volumetric Multi-View Stereo Reconstruction of Manifold Surfaces Based on Dual Graph Embedding," *Proc. Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 503-510, 2006.

[14] V. Kolmogorov and R. Zabih, "Multi-Camera Scene Reconstruction via Graph-Cuts," *Proc. European Conf. Computer Vision,* vol. 3, pp. 82-96, 2002.

[15] V. Kolmogorov and R. Zabih, "What Energy Functions Can Be Minimized via Graph Cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 2, pp. 147-159, Feb. 2004.

[16] K.N. Kutulakos and S.M. Seitz, "A Theory of Shape by Space Carving," *Int'l J. Computer Vision,* vol. 38, no. 3, pp. 199-218, 2000.

[17] V. Lempitsky, Y. Boykov, and D. Ivanov, "Oriented Visibility for Multiview Reconstruction," *Proc. European Conf. Computer Vision,* vol. 3, pp. 226-238, 2006.

[18] S. Osher and J. Sethian, "Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Equations," *J. Computer Physics,* vol. 79, pp. 12-49, 1988.

[19] E. Parzen, "On Estimation of a Probability Density Function and Mode," *Ann. Math. Statistics,* vol. 33, pp. 1065-1076, 1962.

[20] J.-P. Pons, R. Keriven, and O. Faugeras, "Multi-View Stereo Reconstruction and Scene Flow Estimation with a Global Image-Based Matching Score," *Int'l J. Computer Vision,* vol. 72, no. 2, pp. 179-193, 2007.

[21] S. Roy and I.J. Cox, "A Maximum-Flow Formulation of the N-Camera Stereo Correspondence Problem," *Proc. Int'l Conf. Computer Vision,* pp. 735-743, 1998.

[22] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int'l J. Computer Vision,* vol. 47, nos. 1-3, pp. 7-42, 2002.

[23] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," *Proc. Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 519-528, 2006.

[24] S. Sinha and M. Pollefeys, "Multi-View Reconstruction Using Photo-Consistency and Exact Silhouette Constraints: A Maximum-Flow Formulation," *Proc. Int'l Conf. Computer Vision,* vol. 1, pp. 349-356, 2005.

[25] J. Starck, A. Hilton, and G. Miller, "Volumetric Stereo with Silhouette and Feature Constraints," *Proc. British Machine Vision Conf.,* vol. 3, pp. 1189-1198, 2006.

[26] C. Strecha, R. Tuytelaars, and L. Van Gool, "Dense Matching of Multiple Wide-Baseline Views," *Proc. Int'l Conf. Computer Vision,* pp. 1194-1201, 2003.

[27] J. Sun, H.-Y. Shum, and N.-N. Zheng, "Stereo Matching Using Belief Propagation," *Proc. European Conf. Computer Vision,* pp. 510-524, 2002.

[28] S. Tran and L. Davis, "3D Surface Reconstruction Using Graph Cuts with Surface Constraints," *Proc. European Conf. Computer Vision,* vol. 2, pp. 218-231, 2006.

[29] G. Vogiatzis, P.H.S. Torr, and R. Cipolla, "Multi-View Stereo via Volumetric Graph-Cuts," *Proc. Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 391-398, 2005.