A Maximum-Flow Approach to the Volumetric Reconstruction Problem

Catherine Proulx, Sébastien Roy
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, Québec, Canada
cproulx@ieee.org, roys@iro.umontreal.ca

Abstract

We present a 3D reconstruction technique based on the maximum-flow formulation. Starting with a set of calibrated images, we globally search for the most probable 3D model given the photoconsistency and the spatial continuity constraints. This search is done radially from the center of the reconstruction volume; therefore imposing a radial topology. The fact that cameras are arbitrarily positioned around the scene presents challenges for managing occlusion, especially when applying global smoothing. We solve this problem by proposing an iterative occlusion management mechanism, and a new way of looking at surface smoothing and discontinuities that takes photoconsistency into account. Experiments show that our method is relatively fast and robust when dealing with simple objects, even in noisy conditions.

1 Introduction

The problem of passive 3D reconstruction – building a 3D model from a series of calibrated images of a scene – is of great interest in the context of 3D modeling. In theory, this should simply be an extension of the classic stereoscopy problem, which uses the principle of parallax to find the depth of the elements of a scene. In practice, this extension from stereoscopy to a more general camera setup is far from trivial because of the large discrepancy between views, a problem unknown to standard stereo. One obvious stereo-based approach to full volumetric reconstruction is to use two cameras at a time to construct several stereoscopic depth maps and to merge them in a single model. However, this merging process is difficult, especially when the depth maps are noisy. In this paper, our goal was to get rid of this step and go straight from the color matching function to a full 3D geometry. In order to do this, we chose to expand a successful stereoscopy approach, the graph-based energy minimization technique, to the volumetric problem and to propose innovative solutions to the occlusion and smoothing problems.

1.1 Energy-based methods in $2\frac{1}{2}$ D stereoscopy

Energy-based approaches have been widely used in classic $2\frac{1}{2}$ D stereoscopy. The basic idea is to define an energy functional which will be minimal for a 3D reconstruction that best satisfies the matching criteria. These techniques are especially suited to multiple

camera problems because they can easily be expanded to accommodate extra terms or configurations. However, this flexibility is offset by the difficulty in finding a minimum: the more complex the energy function, the harder it is to minimimize it.

The energy function is generally composed of two terms: a matching term which measures how well the solution represents the input images, and a smoothing term which encourages spatial continuity. Other terms can be added to represent additional contraints, at the expense of computational complexity. Different mathematical approaches can be used to solve the energy function. The most generic method is simulated annealing [1], but this method has not been widely used because it tends to be very slow and does not converge to a global minimum. Graph-based methods like belief propagation [11], graph cuts [1] and maximum flow [10] techniques have been proven more efficient and reliable for finding a solution to the stereo problem. These methods are particularly well-suited for problems featuring local dependencies, such as spatial continuity, for instance the stereo problem. On the other hand, they do not handle occlusion very well because of its long-range interactions and its dependence on 3D geometry.

1.2 Photoconsistency-based volumetric reconstruction techniques

The standard in photoconsistency-based volumetric reconstruction is the space carving technique introduced by Kutulakos and Seitz [7]. The basic idea is to scan the reconstruction volume along specific axes and to determine the opacity and the visibility of the voxels following a strict order. Only cameras located behind the current position of a scanning plane can be used in the cost function; the others are assumed to be occluded. Voxels for which the cost function is higher than a certain threshold will be ”carved”, revealing voxels behind them which can then be evaluated similarly. No local smoothing is imposed on the surface, so this method tends to yield disconnected solutions, especially when noise is present. Furthermore, the choice of a suitable threshold is a challenge in itself. In addition, methods based on space carving methods are greedy in nature: the decision to carve a voxel is final and it can only be based on the matching cost of the voxel under study.

Despite these pitfalls, approaches based on space carving have been widely used in 3D reconstruction and several improvements to the basic method have been proposed. The *generalized voxel coloring* technique [2] proposes a different camera management strategy that allows more cameras to be used than in classic space carving. In *approximate n-view stereo*, Kutulakos [6] introduces a more robust cost function, that takes into account not only the projected pixels but their neighbours as well, This refinement makes the space carving method much more resistant to image and calibration noise. Another possible improvement is to reduce the calculation time and memory expense by using an octree representation for the reconstruction volume [9].

1.3 Hybrid techniques

Some hybrid techniques have been proposed to combine the advantages of energy-based approaches and the occlusion management of space carving. Kolmogorov *et al.* [5] model the reconstruction problem as an energy minimization function which contains three terms: the usual matching and smoothing terms, and an additional visibility term that makes the cost of a solution infinite if it is inconsistent with the visibility data.

Unfortunately, the resulting cost function is highly discontinuous and cannot be solved without introducing several simplifications. The visibility modeling seems applicable to stereoscopic scenes with minimal occlusion, but it is unlikely that it could be successfully extended to the fully volumetric problem [3]. Vogiatzis *et al.* [12] start from an initial surface and deform it radially to reconstruct their final model. The optimal displacement for each surface element is computed with the belief propagation algorithm [13], allowing the introduction of a smoothing constraint. Unfortunately, in their approach, occlusion can only be estimated from the initial surface – which means that the results are highly dependent on the quality of that solution – and only synthetic volumetric results have been provided so far.

2 Algorithm

2.1 Reconstruction volume

Reconstruction volumes can be divided into two categories. On one side, we find the purely volumetric approach where the reconstruction space is divided into evenly spaced cubic voxels. The reconstruction problem can then be solved by assigning a state - filled or empty - to each voxel. This is an elegant but costly approach since every unit must be evaluated independently in order to find a solution. On the other side, height-field approaches try to reduce the search space by using an initial surface and finding the displacement normal to that surface that best represents the scene. In classic stereoscopy, this surface is a plane but as Vogiatzis *et al.* [12] have shown, we can generalize the method to arbitrary surfaces. The pitfall of this approach is that the displacement from the initial surface must be very small for two reasons: first, the visibility calculation depends on this initial surface, so large displacements will create occlusions not managed by the visibility algorithm, and second, auto-intersection problems may occur when the displacement is significant with respect to the curvature of the surface (figure 1).

We propose an ”onion-shaped” reconstruction volume instead, i.e. a spherical volume divided into layers of uniform thickness (figure 2). Each layer is tesselated quasi-uniformly and a voxel is formed around each vertex of the resulting mesh. Excluding the scale factor, the tesselation is the same for each layer. This means that we can define a visiting order from the exterior to the core of the volume that passes through the n^{th} voxel of each layer. We can also define neighbourhood relationships between voxels of a layer.

The volume we use is a hybrid between both techniques: it still relies on some initial knowledge of the scene but allows for large variations from that initial topology. As in volumetric reconstruction, the whole search space is discretized in small volumetric units. However, these voxels are not cubic nor uniformly sized. They are aligned along vectors coming out of a central point, which means that our volume can be seen as a specific case of a height-field on a spherical surface. The use of a very specific surface means that we avoid auto-intersection problems even with very large displacements from the initial surface. On the other hand, we are limited to objects featuring a radial topology [1].

[1] An object is defined as having a radial topology if there is a point c, called the center, such that all the rays r starting from it cross the object’s surface only once.

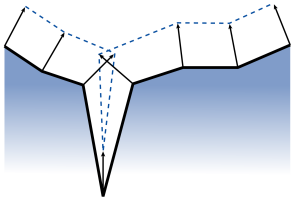


Figure 1: The auto-intersection problem in height-field approaches with a large displacement from the initial surface.

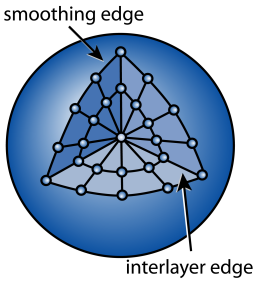


Figure 2: Our reconstruction volume: spherical layers are embedded in one another, and a voxel is formed around each vertex of the tesselated sphere.

2.2 Energy function and graph representation

We propose to solve the reconstruction problem by finding the minimum of an energy function of the form $E(f) = E_{data}(f) + E_{smoothness}(f)$, where $E_{data}(f)$ which denotes the "matching cost" and $E_{smoothness}(f)$ the "smoothing cost". Our strategy is to use a maximum flow/minimum cut algorithm [10], although other energy minimization methods could be used. This is done by building an undirected graph composed of a source s, a sink t and a network of nodes and edges such that any s-t cut of the graph represents a valid closed reconstruction surface. Moreover, the capacity of the edges is chosen so that finding the cut with the minimum total edge cost corresponds to finding the surface minimizing our energy function. In our case, we form the graph by associating a node to each voxel of the reconstruction volume, with smoothing edges between voxels of a layer, and interlayer edges along the paths to the core where the sink t is located. By pushing flow from the exterior of the spherical graph to its core, we find the area where the flow saturates, which corresponds to the best reconstruction surface. More formally, given a reconstruction volume with D layers of N vertices per layer, the graph $G(V,E)$ consists in:

$$V = \{(a,d) \mid 1 \leq a \leq N, 1 \leq d \leq D\} \cup \{s,t\} \tag{1}$$

$$E = E_{smoothness} \cup E_{matching} \cup E_{source} \cup E_{sink} \quad \text{where}$$

$$E_{smoothness} = \{((a,d)(b,d)) \mid 1 \leq a \leq N, b \in \mathcal{N}_a, 1 \leq d \leq D\}$$

$$E_{matching} = \{((a,d)(a,d+1)) \mid 1 \leq a \leq N, 1 \leq d \leq D-1\}$$

$$E_{source} = \{(s,(a,D)) \mid 1 \leq a \leq N\}, \quad E_{sink} = \{((a,1),t) \mid 1 \leq a \leq N\}$$

$$\mathcal{N}_a \quad \text{is the layer neighbourhood of vertex } a. \tag{2}$$

2.3 Spatial smoothing in a volumetric context

The next step of our algorithm is to define the capacity of the edges of the graph. In graph-based $2\,1/2$ D stereoscopy, interlayer edges implement the matching cost while smoothing edges implement the bias for spatial continuity. It is typical to consider that this bias is identical everywhere and to use a constant value for the capacity of the smoothing edges. Some researchers have experimented with techniques which modulate the capacity

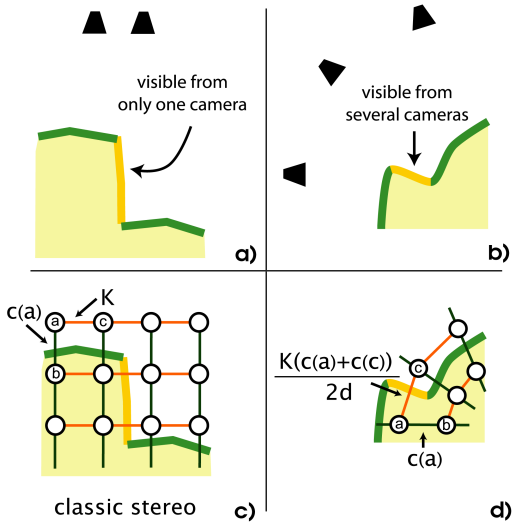


Figure 3: Capacity of the interlayer and smoothing edges.

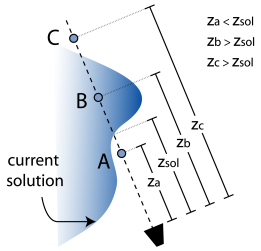


Figure 4: Measure of occlusion.

of those edges with data from the reference image, for instance the presence of a 2D contour at that location that makes the presence of a spatial discontinuity more likely [1]. However, this kind of method does not generalize well to the volumetric case, where there is no ”reference” image.

We choose a truly volumetric approach instead, which is based on the observation that in our mathematical model, all the edges that are part of the minimum cut are associated with a small segment of the reconstructed surface. In classic stereoscopy, we assume that the final scene is composed of fronto-parallel planes, and that the surface elements that link them across discontinuities, i.e. the surface elements associated with the smoothing edges, are invisible from most cameras (Figure 3a). This hypothesis holds true as long as the distance between the cameras is small and the number of cameras is limited. However, in a typical volumetric scene, the cameras are arranged in such a way that every element of the surface can be seen properly from several points of view, even the elements associated with smoothing edges. Therefore, it makes sense to incorporate the available photoconsistency information to make a good decision concerning these edges, and not penalizing discontinuities consistent with the color data. (Figure 3b).

Let us assume that a cost function c evaluating the photoconsistency of voxels is defined, taking occlusion into account, c being small for good matching costs. We propose such a cost function in the next section, but the solving method can be used with any cost function respecting the previous criteria. We will directly use the value of the cost function at a voxel for the capacity of the interlayer edge which emerges from it. The capacity of the smoothing edges will depend on the average cost of the two voxels they link, multiplied by a factor representing the amount of smoothness of the reconstruction. Since the radial distance between the layers is constant, it is not necessary to model the distance between voxels in the capacity of interlayer edges. The smoothing edges, on the other hand, must take it into account: we assume that closer voxels should be more

interrelated than distant ones, so we divide the base capacity of the smoothing edges by the distance between them. Hence, the final edge capacity is (figure 3c and d):

$$capacity(u,v) = \begin{cases} \infty & \text{if } (u,v) \in E_{source} \\ c(u) & \text{if } (u,v) \in E_{sink} \\ c(u) & \text{if } (u,v) \in E_{matching} \\ \frac{K}{2\|p_v - p_u\|} \frac{c(u)+c(v)}{2} & \text{if } (u,v) \in E_{smoothing} \end{cases} \quad (3)$$

where $c(x)$ is the cost of the voxel associated with node x, K is a factor which determines the amount of smoothing, p_x is the 3D position of the center of voxel x and $\|\|$ is the Euclidian norm.

2.4 Cost function and occlusion model

It is a well-known fact that in volumetric reconstruction, occlusion is omnipresent and must be modeled explicitly [5]. We have seen earlier that approaches based on space carving deal with visibility sequentially. If we want to solve the problem globally, a different occlusion management approach will be necessary. One alternative is to add a visibility term inside the cost function, as done in Kolmogorov *et al.*, but this increases the complexity of the energy function significantly and makes the problem almost intractable. A more practical option is to consider the problem iteratively, using a previously obtained surface as an approximate solution to determine visibility [8, 4]. However, this approach leads to a chicken-and-egg dilemma: we need a first solution to determine the visibility, but we need the visibility to reconstruct this first solution.

This forces us to rely on the simplifying assumption that there is initially no occlusion. We have observed that even though it is an incorrect assumption, it is sufficient to reconstruct an initial coarse reconstruction that roughly represents the scene. We then use this coarse reconstruction to evaluate the visibility and to eliminate a subset of the cameras which we judge to be occluded. Occlusion is measured by rendering the current solution from a camera's point of view and using the depth buffer to evaluate the depth of the surface z_{sol} (figure 4). If this surface is placed in front of the voxel we are examining, i.e. $z_{sol} < z_{voxel}$, we assume that this camera is in occlusion with respect to this voxel and that the magnitude of this occlusion is proportional to the distance between the voxel and the surface $|z_{sol} - z_{voxel}|$.

Of course, our confidence in the initial solution is limited, so we must be conservative in our camera elimination. In order to guarantee convergence in a set number of iterations, it is also suitable to make this process permanent: once a camera has been eliminated for a given voxel, it will never contribute again to its cost function. After a certain number of cameras (one, in our tests) have been eliminated, we recreate a more refined solution with the energy minimization algorithm. In theory, we could iterate like this until no more cameras can be eliminated, or only two cameras remain. However, empirical results have shown that five to ten iterations are sufficient to produce a good reconstruction. We complement this iterative mechanism with a robust measure of photoconsistency that helps produce a reasonably good reconstruction even when visibility information is erroneous. Each voxel is projected into all the cameras from which it is visible; all the intensity values are then compared two by two and the minimum color distance between two values is used as our cost function. Explicitly:

$$c(v) = \min_{i \neq j} \left| I_i(v) - I_j(v) \right| \tag{4}$$

where $c(v)$ is the cost of voxel v and $I_n(v)$ is the average color of voxel v when projected into image n.

Naturally, this function is too simplistic to produce a good reconstruction: it yields very rough data with a lot of uncertainty, especially in the first few iterations when little or no camera visibility information is used. However, the maximum flow algorithm deals very well with this kind of data, yielding a coarse but smooth and well-formed solution. Also, this cost function generally deals well with specularities, though it is based on the assumption that the object surface is lambertian, because specularities will be treated as outlier samples. One the other hand, the method is very sensitive to uniform backgrounds, therefore it is advisable to design a scene with highly textured backgrounds or to pre-process the images with a random background. The latter approach was used in this paper.

3 Experimental results

The results were generated at two different resolutions: the low resolution volume is composed of 50 layers of 2 562 vertices each, for a total of 128 100 voxels, and the high resolution volume of 40 layers of 10 242 vertices each, for a total of 409 680 voxels. At our highest resolution, the unoptimized calculation typically required 30 to 40 minutes (on a 1.60 GHz Pentium M processor), and most of this time was spent on the projection of the voxels and the computation of the cost function. We observed that the computation time is linear with respect to the number of voxels.

3.1 Synthetic scenes

A first series of tests was conducted with synthetic images generated in OpenGL. These images represent a best-case scenario for our algorithm: the objects have a simple geometry that features a radial topology for the most part, their surface is perfectly lambertian and well-textured, and the cameras are placed fairly evenly on the surface of a sphere that encompasses the reconstruction volume. Two such results are presented here. Figure 5 presents a high-resolution reconstruction of a relatively complex model: we can see that the details of the surface are well recovered, except in the regions where the radial topology constraint is not respected (mostly on the character's pacifier). There is some sampling noise on the surface, but it should be easily reduced by increasing the resolution of the surface. Figure 6 shows the impact of the edge capacity model described in section 2.3 on a low-resolution reconstruction. We observe that the results on the right, obtained with our new smoothing approach, present much more accurate edges while smooth surfaces have approximately the same amount of noise. This is due to the fact that the smoothing is non-uniform, and that the discontinuities on the edges of the model are consistent with the color data, and therefore not penalized the way noise peaks would be. In both cases, we used a smoothing factor of 0.005. In fact, we observed that this smoothing factor is independent from the resolution of the scene, and that the same factor yields good results for a wide variety of scenes.

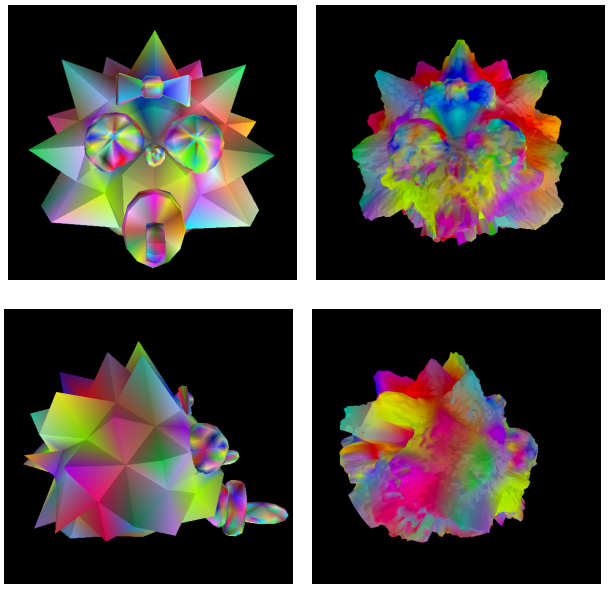
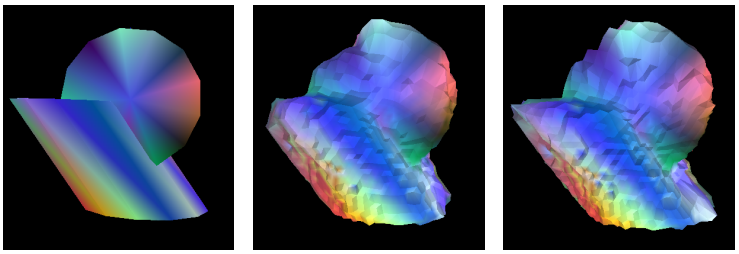


Figure 5: Synthetic results: two views of the *maggie* model with their corresponding high resolution reconstruction.



(a) (b) (c)

Figure 6: Impact of the capacity of the smoothing arcs: a) original view of the *canon* model b) reconstruction with the traditional approach in which only distance between nodes influences the smoothing c) low resolution reconstruction with our new formulation for the smoothing arcs capacity, which takes photoconsistency into account. The resolution and average capacity of the smoothing arcs is the same for both reconstructions.

3.2 Real images

Our second series of tests were conducted with digital photographs taken under real-life conditions [2]. Sixteen images were taken with a turntable setup under uniform lighting. All cameras were located slightly below the center of gravity of the model, in a circular configuration, and they were calibrated with Tsai's algorithm. The calibration presents reprojection errors of ± 0.5 pixels. As stated earlier, images were manually segmented and a random color was applied to their background. We used the same volume and smoothing

[2] Images curtesy of Kyros Kutulakos, from the University of Toronto

Figure 7: Two views of the *gargoyle* model with corresponding reconstruction. We rendered the reconstructed model by associating to each vertex the average projected color of the corresponding voxel.

factor as for the *maggie* synthetic sequence. Results are shown in figure 7. Note that due to the radial topology constraint, only the head of the gargoyle was reconstructed.

4 Conclusion

In this paper, we have proposed a new energy-based formulation for the volumetric reconstruction problem based on the retrieval of a minimum cut in a spherical graph. We presented a new strategy for dealing with surface discontinuities that allows uneven smoothing depending on the photoconsistency of the voxels that form this discontinuity. Thanks to this strategy, we are able to distinguish accurately between noise peaks and actual surface discontinuities, and to prevent over-smoothing. We also presented a new way of dealing with occlusion, inspired in part by the works of Nakamura *et al.* [8] and Kang *et al.* [4] in classic stereo, which refines the visibility information iteratively.

The main limitation of our algorithm is clearly the radial topology constraint. A few avenues have been explored to sidestep this problem – namely the use of deformable reconstruction volumes and multi-sink graphs – but these options will have to be studied more in depth before we can achieve a truly general approach. A more formal analysis of our algorithm and of its convergence conditions would also be necessary. Still, we consider that this paper confirms that energy-based approaches are a promising alternative to the space carving algorithms in volumetric reconstruction.

5 Acknowledgements

This research was supported financially by the Natural Sciences and Engineering Research Council of Canada (NSERC). Special thanks to Kyros Kutulakos for his images.

References

[1] Y. Boykov, O. Veksler, et al. Fast approximate energy minimization via graph cuts. In *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 377–384. 1999.

[2] W. B. Culbertson, T. Malzbender, et al. Generalized voxel coloring. In *Lecture Notes in Computer Science*, vol. 1883, pp. 100–115. 1999.

[3] M.-A. Drouin, M. Trudeau, et al. Go-consistency for wide multi-camera stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2005.

[4] S. B. Kang, R. Szeliski, et al. Handling occlusions in dense multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 103–110. 2001.

[5] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *European Conference on Computer Vision*, pp. 82–96. 2002.

[6] K. N. Kutulakos. Approximate n-view stereo. In *European Conference on Computer Vision*, pp. 67–83. 2000.

[7] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. Tech. Rep. TR692, Comp. Science Department, U. Rochester, 1998.

[8] Y. Nakamura, T. Matsuura, et al. Occlusion detectable stereo: Occlusion patterns in camera matrix. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1996.

[9] A. C. Prock and C. R. Dyer. Towards real-time voxel coloring. In *Proc. Image Understanding Workshop*, pp. 315–321. 1998.

[10] S. Roy and I. J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 492–502. 1998.

[11] J. Sun, N.-N. Zheng, et al. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800, 2003.

[12] G. Vogiatzis, P. Torr, et al. Reconstructing relief surfaces. In *British Machine Vision Conference*, pp. 117–126. 2004.

[13] J. S. Yedidia, W. T. Freeman, et al. *Understanding Belief Propagation and Its Generalizations*, chap. 8. Morgan Kaufmann, 2002.