Robust Variational Segmentation of 3D Objects from Multiple Views

Kalin Kolev, Thomas Brox, and Daniel Cremers

CVPR Group, University of Bonn
Römerstr. 164, 53117 Bonn, Germany
{kolev, brox, dcremers}@cs.uni-bonn.de

Abstract. We propose a probabilistic formulation of 3D segmentation given a series of images from calibrated cameras. Instead of segmenting each image separately in order to build a 3D surface consistent with these segmentations, we compute the most probable surface that gives rise to the images. Additionally, our method can reconstruct the mean intensity and variance of the extracted object and background. Although it is designed for scenes, where the objects can be distinguished visually from the background (i.e. images of piecewise homogeneous regions), the proposed algorithm can also cope with noisy data. We carry out the numerical implementation in the level set framework. Our experiments on synthetic data sets reveal favorable results compared to state-of-the-art methods, in particular in terms of robustness to noise and initialization.

1 Introduction

Recovering the spatial structure of a scene from multiple views is one of the oldest and most fundamental problems in computer vision with many applications in computer graphics, robot navigation, object recognition, and tracking. The literature on 3D reconstruction could be divided into four major classes: shape from stereo, shading, texture, and silhouettes.

Stereovision requires to match points from different images that correspond to the same point in the scene. The earliest algorithms that incorporate a large number of views use carving techniques to obtain a volumetric representation of the scene assuming Lambertian properties of the objects [18,10]. The space carving framework suffers from several limitations. Once a voxel is carved away, it cannot be recovered. Moreover, if one voxel is removed in error, further voxels can be erroneously removed in a cascade effect. These limitations are partially alleviated by the probabilistic space carving method [1]. Others have suggested to guide a deformable surface model by a measure based on local correspondences toward a steady state [6,5]. All these methods require a textured surface in order to match points.

Shape from shading methods, on the other hand, are mainly designed for homogeneous objects [8,9]. They are based on the diffusing properties of Lambertian surfaces and aim at reconstructing the object shape from light reflectance.

K. Franke et al. (Eds.): DAGM 2006, LNCS 4174, pp. 688–697, 2006.

A difficulty of this concept is the requirement of a known illumination model or the necessity to estimate illumination together with the shape.

A similar problem appears with texture-based methods [12]. They need a known texture pattern in order to reconstruct a 3D surface by means of its distortion in the image.

In case of sparsely textured objects, which are known challenges to stereo- and texture-based techniques, silhouettes exhibit the dominant image feature. The algorithm presented in this paper belongs to this type of silhouette-based techniques. Such methods usually try to estimate the *visual hull* of the observed objects. The visual hull of an object is defined as the maximal shape that yields the same silhouette as the observed object [11]. The earliest attempts use a volumetric representation of the scene and are referred to as *volume intersection* techniques in the literature. That is, the space is discretized by a fixed voxel grid and each voxel is labeled as opaque or transparent. An early paper reporting a volumetric representation of the visual hull is due to Martin and Aggarwal [13]. They segment the input images in advance by a simple intensity thresholding and then back-project the estimated silhouettes to a surface representation. Since then, silhouettes have been used in many different algorithms. Octree-based representations have been employed by [15,19,7], and in [17] the authors presented a Hough-like voting scheme that back-projects image features into a volumetric space. In addition to volumetric approaches, some surface-based ones have been presented. In [3] and [20] apparent contours are used to reconstruct a 3D shape. Although the authors obtain better results, the reconstruction works only locally.

Yezzi and Soatto recently proposed *stereoscopic segmentation* as a variational framework for global 3D region segmentation from a collection of images of a scene [21]. They couple the segmentations of each image through the evolution of a single 3D surface rather than separate 2D contours, which makes their method robust to erroneous camera calibration. Upon a closer look, it turns out that stereoscopic segmentation has certain limitations. Its main drawback is the definition of the energy in the image domain that results in a very local evolution. Consequently, it needs an accurate initialization in order to capture the correct object topology. In addition, the algorithm is prone to noise as the strictly local surface evolution is mainly determined by single camera observations.

In this paper, we propose a probabilistic Bayesian formulation of 3D reconstruction which aims at estimating the most likely 3D shape given the observed images. In contrast to stereoscopic segmentation, this yields a more global evolution that makes better use of the available information from multiple cameras. As a consequence, our method has a larger radius of convergence and is more robust to noise than previous techniques.

Paper organization. In the next section, the probabilistic framework of the proposed method is presented and discussed. A variational formulation and a respective level set implementation are developed in Section 3. In Section 4 we show experimental results. Finally, we provide a conclusion in Section 5.

2 Probabilistic Volume Intersection

2.1 Bayesian Inference

Let V be a discretized volume and $I_1, \ldots, I_n : \Omega \mapsto \mathbb{R}$ a collection of calibrated input images with perspective projections $\pi_1, \ldots, \pi_n$. Given the set of images, we are looking for the most probable surface $\hat{S}$ that gives rise to these images, that is

$$\hat{S} = \arg\max_{S \in \Lambda} P(S \mid \{I_1, \ldots, I_n\}), \tag{1}$$

where Λ is the set of all closed surfaces lying inside of the volume V. By means of the Bayes formula we obtain (omitting the normalization constant):

$$P(S \mid \{I_1, \ldots, I_n\}) \propto P(\{I_1, \ldots, I_n\} \mid S) \cdot P(S). \tag{2}$$

Assuming that all voxels are independent leads to

$$P(S \mid \{I_1, \ldots, I_n\}) \propto \left[\prod_{x_{ijk} \in V} P(\, \{I_l(\pi_l(x_{ijk}))\}_{l=1,\ldots,n} \mid S) \right]^{dx} \cdot P(S), \tag{3}$$

where dx denotes the discretization step. The exponent dx is introduced to ensure the correct continuum limit. The resulting expression is then invariant to refinement of the grid.

According to a certain surface estimate S, the voxels can be divided into two classes: lying inside an object or belonging to the background. Hence, the volume V can be expressed as $V = R^S_{obj} \cup R^S_{bck}$. Considering this partitioning, we can proceed with

$$P(S \mid \{I_1, \ldots, I_n\}) \propto \left[\prod_{x_{ijk} \in R^S_{obj}} P(\, \{I_l(\pi_l(x_{ijk}))\}_{l=1,\ldots,n} \mid x_{ijk} \in R^S_{obj}) \right]^{dx} \cdot$$
$$\left[\prod_{x_{ijk} \in R^S_{bck}} P(\, \{I_l(\pi_l(x_{ijk}))\}_{l=1,\ldots,n} \mid x_{ijk} \in R^S_{bck}) \right]^{dx} \cdot P(S).$$

To simplify the notation, we denote

$$\begin{aligned} P_{obj}(x) &:= P(\, \{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{obj}) \\ P_{bck}(x) &:= P(\, \{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{bck}) \end{aligned} \tag{4}$$

for $x \in V$ (see fig. 1) and come to the following expression

$$\hat{S} = \arg\max_{S \in \Lambda} \left[\prod_{x_{ijk} \in R^S_{obj}} P_{obj}(x_{ijk}) \cdot \prod_{x_{ijk} \in R^S_{bck}} P_{bck}(x_{ijk}) \right]^{dx} \cdot P(S). \tag{5}$$

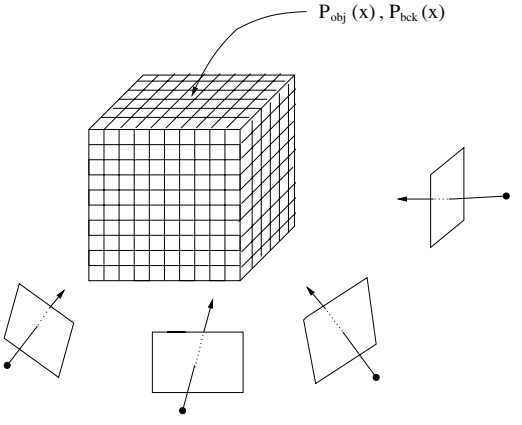

Fig. 1. Volume representation. Two probabilities P_{obj}, P_{bck} are assigned to each voxel for membership to one of the objects and background, respectively.

2.2 Joint Probabilities

In order to compute the joint probabilities $P(\ \{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{obj}\ )$ and $P(\ \{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{bck}\ )$, we have to combine information from different images. This could be achieved by assuming independence of the image observations yielding

$$\begin{aligned} P_{obj}(x) &= \prod_{i=1}^{n} P(I_i(\pi_i(x)) \mid x \in R^S_{obj}) \\ P_{bck}(x) &= 1 - \prod_{i=1}^{n} \left[1 - P(I_i(\pi_i(x)) \mid x \in R^S_{bck})\right]. \end{aligned} \tag{6}$$

Note the asymmetry in these expressions. The probability of a voxel being part of the foreground is equal to the probability that all cameras observe this voxel as foreground, whereas the probability of background membership describes the probability of at least one camera seeing background. However, this model has some disadvantages. In case of noisy images $0 < P(\ I_i(\pi_i(x)) \mid x \in R^S_{obj}\ ) < 1$ and $0 < P(\ I_i(\pi_i(x)) \mid x \in R^S_{bck}\ ) < 1$, in general. Hence, for $n \to \infty$ the joint probability $P(\ \{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{obj}\ )$ will converge to 0 and $P(\ \{I_l(\pi_l(x))\}_{l=1,\ldots,n} \mid x \in R^S_{obj}\ )$ to 1. To dispose this bias for increasing number of cameras, we have to take the dependency of the observations into account. In our model we used the geometric mean of the single probabilities:

$$\begin{aligned} P_{obj}(x) &= \sqrt[n]{\prod_{i=1}^{n} P(I_i(\pi_i(x)) \mid x \in R^S_{obj})} \\ P_{bck}(x) &= 1 - \sqrt[n]{\prod_{i=1}^{n} \left[1 - P(I_i(\pi_i(x)) \mid x \in R^S_{bck})\right]}. \end{aligned} \tag{7}$$

They are modeled by Gaussian densities

$$\begin{aligned} P(I_i(\pi_i(x)) \mid x \in R^S_{obj}) &= \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{-\frac{(I_i(\pi_i(x))-\mu_{obj})^2}{2\sigma^2}} \\ P(I_i(\pi_i(x)) \mid x \in R^S_{bck}) &= \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{-\frac{(I_i(\pi_i(x))-\mu_{bck})^2}{2\sigma^2}}, \end{aligned} \tag{8}$$

where μ_{obj}, μ_{bck} denote the mean intensities of object/background and σ is the respective standard deviation. We update these values in the course of evolution by projecting the current surface estimate onto the images as described in [21]. The standard deviation σ is set to the maximum of the deviations of the object and background regions. Alternatively, above probabilities could be modeled with two separate standard deviations. However, in our experiments the proposed model resulted in a faster convergence.

3 Variational Framework

3.1 Variational Formulation

In this section we will convert the maximum a-posteriori estimation into an energy minimization problem. Applying the negative logarithm to (5) yields in a continuous formulation the following functional:

$$E(S) = -\int_{R^S_{obj}} \log P_{obj}(x)\, dx - \int_{R^S_{bck}} \log P_{bck}(x)\, dx - \log P(S). \tag{9}$$

Minimizing this energy functional is equivalent to maximizing the total a-posteriori probability of all voxel assignments. The first two terms are related to the external energy and measure the discrepancy between observed images and images predicted by the model. The last term exhibits the internal energy and describes the surface shape, thus allowing incorporation of prior knowledge on the geometry. Note that the functional also incorporates the intensity means and standard deviation, which are defined by the surface S. Since the unknowns, surface and radiances, live in an infinite-dimensional space (there exist multiple solutions S that explain the observed images), we need to impose regularization in order to make the minimization problem well-posed. This can be achieved by setting

$$P(S) = e^{-\nu|S|}, \tag{10}$$

where ν is a weighting constant and $|S|$ denotes the surface area. Inserting this expression into the above functional yields

$$E(S) = -\int_{R^S_{obj}} \log P_{obj}(x)\, dx - \int_{R^S_{bck}} \log P_{bck}(x)\, dx + \nu|S|. \tag{11}$$

In order to reconstruct the smoothest surface consistent with the images, we omit the data fidelity terms for points, which are visible from neither of the cameras. This is not restrictive, since no data is available for such points.

3.2 Level Set Implementation

The numerical implementation of the proposed energy functional (11) has been carried out within the level set framework [4,14] due to its stability and ability to handle topological changes automatically. In level set methods, the surface is implicitly represented by a function $\phi : V \mapsto \mathbb{R}$, whose values are the distances from the surface, and the interior and exterior of the surface are defined by $\phi(x) < 0$ and $\phi(x) \geq 0$, respectively. Hence, we can use the Heaviside function

$$H(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

to access these two regions. Expressing the energy functional (11) with respect to the level set function ϕ yields

$$\begin{aligned} E(\phi) = &- \int_V [\log P_{bck}(x) H(\phi(x)) + \log P_{obj}(x)(1 - H(\phi(x)))] \, dx \\ &+ \nu \int_V |\nabla H(\phi(x))| dx. \end{aligned} \tag{13}$$

This formulation has some nice properties. First, its Euler-Lagrange equations are easy to compute since the implicit function ϕ occurs as an argument. Second, it leads to a stable volume-based surface flow. A similar energy functional was used in [2,16] for image segmentation purposes. The Euler-Lagrange equations of (13) read

$$\frac{\partial \phi(x)}{\partial t} = \delta(\phi(x)) \cdot [\log P_{bck}(x) - \log P_{obj}(x)] + \nu \delta(\phi(x)) \cdot div \left(\frac{\nabla \phi(x)}{|\nabla \phi(x)|} \right), \tag{14}$$

where $\delta(\cdot)$ denotes the Dirac function

$$\delta(z) = \frac{d}{dz} H(z). \tag{15}$$

In practice, smoothed versions of $H(\cdot)$ and $\delta(\cdot)$ have to be applied [2].

4 Experiments

In Fig. 2 we show results obtained with the proposed algorithm applied to 20 noisy images, four of which are depicted in Fig. 2(a). Fig. 2(c) visualizes the final result from multiple viewing directions. Obviously, our method is able to deal with noise as well as lighting effects and leads to an accurate reconstruction of the two balls. In order to emphasize its robustness a reconstruction generated by carving techniques is presented for comparison. For the sake of fairness we added an identical smoothness term in the implementation of the shape carving method. The estimated mean intensities computed by our algorithm were used for segmenting the input images separately and independently. As clearly visible in Fig. 2(d), this approach is susceptible to noise and shading effects, since only

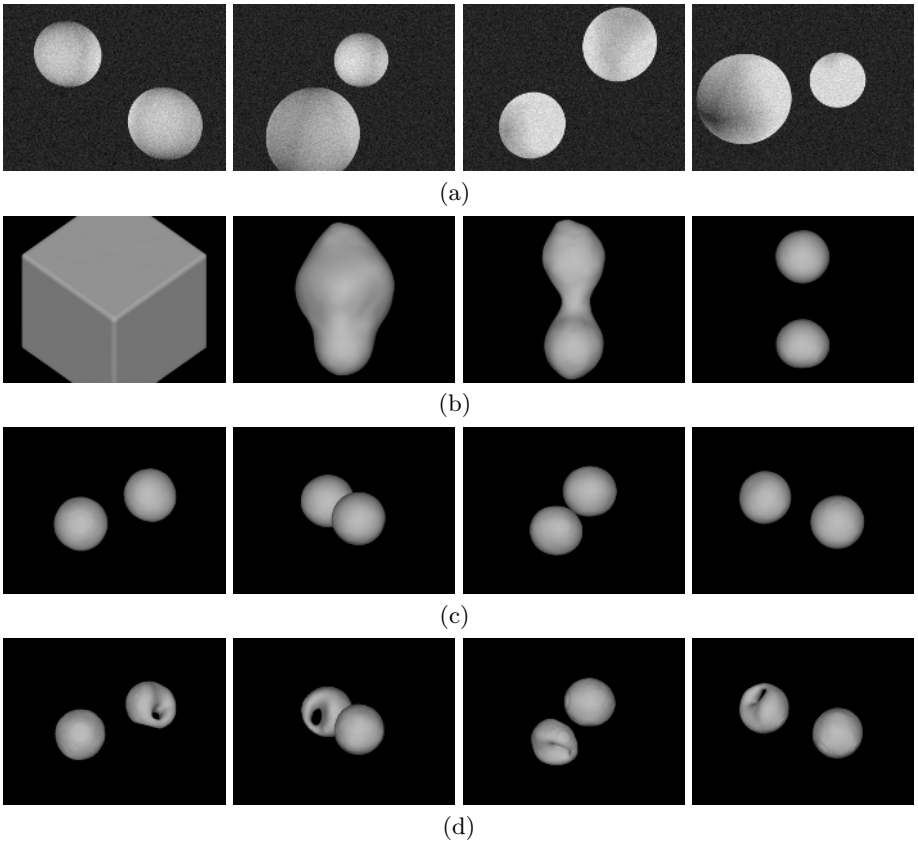

Fig. 2. Reconstruction of two spheres. (a) 4 out of 20 input images disturbed by noise, (b) surface during evolution, (c) reconstructed surface obtained with our probabilistic method, (d) result obtained with carving techniques.

single observations are taken into account for deciding whether a voxel should be carved away or not. In contrast, our method is quite robust to noise due to the averaging effect of integrating data from all views.

Fig. 3 demonstrates the ability of the proposed method to reconstruct complex topologies starting with an arbitrary initialization as opposed to stereoscopic segmentation, which requires an approximation of the real topology, as stated in [21]. The reconstructions of a torus obtained with our method and with stereoscopic segmentation from the same initial surface are depicted in Fig. 3(c) and Fig. 3(d), respectively. Note that, similar to stereoscopic segmentation, our method is bidirectional, i.e., surfaces can evolve inward as well as outward. In addition, our formulation leads to a surface evolution that allows for bigger time steps. In contrast to stereoscopic segmentation, the time step size is only restricted by the smoothness constraint.

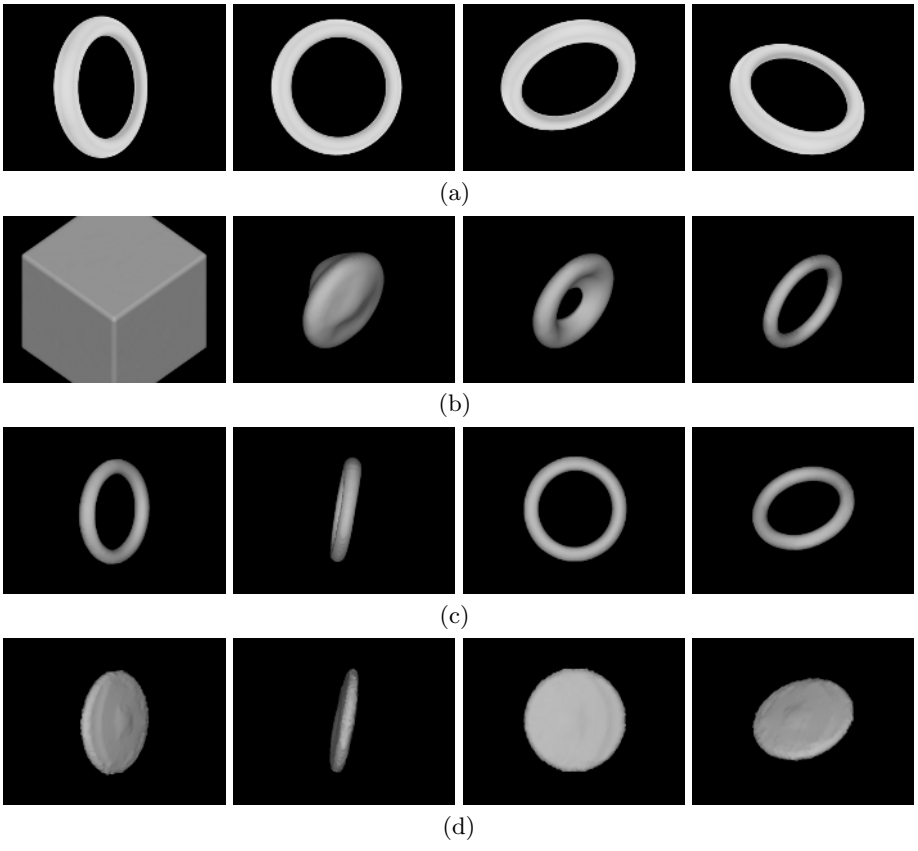

Fig. 3. Reconstruction of a torus. (a) 4 out of 20 input images, (b) surface during evolution, (c) reconstructed surface obtained with our method, (d) result obtained with stereoscopic segmentation [21] from the same initialization.

Finally, Fig. 4 illustrates the behavior of the presented algorithm when applied to a data set that exhibits ambiguous silhouette information. The cameras are arranged in such a way that none of them can see the bottom of the vase. Due to the geometric prior, the lacking information results in the smoothest shape that is photometrically consistent with the data (note the flat bottom and the neck of the vase).

All illustrated results were obtained from 20 images with 640×480 pixels using a C++ implementation running on a Pentium IV with 3.4GHz. All cameras were situated on a bounding sphere enclosing the scene. For a cubic grid of $128 \times 128 \times 128$ the algorithm takes between 20 and 30 minutes to converge, which is about a factor 3 faster than stereoscopic segmentation. Moreover, it can still be substantially accelerated when replacing our preliminary surface projection algorithm by a more sophisticated implementation.

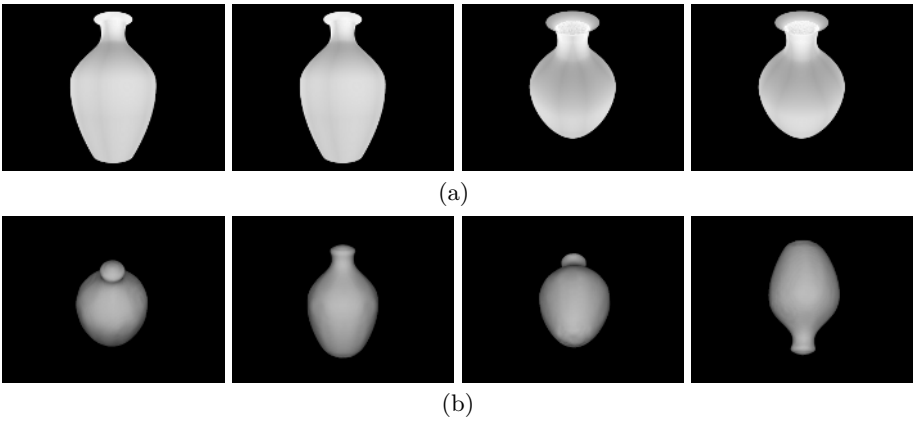

Fig. 4. Reconstruction of a vase. (a) 4 out of 20 input images. Due to the rotational symmetry and the arrangement of the cameras, most images look the same. (b) Reconstructed surface from multiple views.

5 Summary

We have presented a new variational approach to reconstruct smooth shapes from a number of calibrated camera views. The variational formulation is derived from a probabilistic setting via Bayesian inference and uses the level set framework to represent the sought object surface. The mean radiance of object and background are estimated together with the surface. In comparison to previous methods, the probabilistic derivation and formulation of the energy on the volumetric instead of the image domain provides faster convergence and better robustness to noise or other violations of the assumption of constant object radiance. Moreover, the optimization is less prone to accurate initializations and allows to reconstruct more complex topologies. These properties have been confirmed in experimental evaluation. Future work is focused on applications to real data sets.

Acknowledgments

This work was supported by the German Research Foundation, grant #CR-250/1-1.

References

1. A. Broadhurst, T. W. Drummond, and R. Cipolla. A probabilistic framework for space carving. In *Proc. International Conference on Computer Vision*, pages 388–393, July 2001.
2. T. Chan and L. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, Feb. 2001.

3. R. Cipolla and A. Blake. Surface shape from the deformation of apparent contours. *International Journal of Computer Vision*, 9(2):83–112, 1992.
4. A. Dervieux and F. Thomasset. A finite element method for the simulation of Rayleigh–Taylor instability. In R. Rautman, editor, *Approximation Methods for Navier–Stokes Problems*, volume 771 of *Lecture Notes in Mathematics*, pages 145–158. Springer, Berlin, 1979.
5. Y. Duan, L. Yang, H. Qin, and D. Samaras. Shape reconstruction from 3D and 2D data using PDE-based deformable surfaces. In *Proc. European Conference on Computer Vision*, pages 238–251, 2004.
6. O. Faugeras and R. Keriven. Variational principles, surface evolution, PDE's, level set methods, and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):336–344, Mar. 1998.
7. B. Garcia and P. Brunet. 3D reconstruction with projective octrees and epipolar geometry. In *Proc. International Conference on Computer Vision*, pages 1067–1072, January 1998.
8. B. Horn and M. Brooks. *Shape from shading*. MIT Press, 1989.
9. H. Jin, D. Cremers, A. Yezzi, and S. Soatto. Shedding light on stereoscopic segmentation. In L. Davis, editor, *Proc. International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 36–42, Washington, DC, 2004.
10. K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.
11. A. Laurentini. The visual hull concept for visual-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.
12. J. Malik and R. Rosenholtz. A differential method for computing local shape-from-texture for planar and curved surfaces. In *Computer Vision and Pattern Recognition Conference*, pages 267–273, June 1993.
13. W. N. Martin and J. K. Aggarwal. Volumetric descriptions of objects from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):150–158, 1983.
14. S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *Journal of Computational Physics*, 79:12–49, 1988.
15. M. Potmesil. Generating octree models of 3D objects from their silhouettes from a sequence of images. *Computer Vision, Graphics, and Image Processing*, 40(1):1–29, 1987.
16. M. Rousson, T. Brox, and R. Deriche. Active unsupervised texture segmentation on a diffusion based feature space. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 699–704, Madison, WI, June 2003.
17. S. Seitz and C. Dyer. Complete scene structure from four point correspondences. In *Proc. International Conference on Computer Vision*, pages 330–337, June 1995.
18. S. Seitz and C. Dyer. Photorealistic scene reconstruction by voxel coloring. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 1067–1073, June 1997.
19. R. Szeliski. Rapid octree construction from image sequences. *Computer Vision, Graphics, and Image Processing*, 58(1):23–32, 1993.
20. R. Vaillant and O. Faugeras. Using extremal boundaries for 3D object modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):157–173, 1992.
21. A. Yezzi and S. Soatto. Stereoscopic segmentation. *International Journal of Computer Vision*, 53(1):31–43, 2003.