

A study of 3D reconstruction algorithms from multiple views

Hassan Hajjdiab
 Abu Dhabi University,
 Abu Dhabi, UAE,
 hassan.hajjdiab@adu.ac.ae,
<http://www.adu.ac.ae>

Abstract

Constructing a three-dimensional (3D) model of scene from a set of two-dimensional (2D) images is a challenging problem in computer vision and artificial intelligence. Solving this visual reconstruction problem has many applications ranging from medical imaging, robot navigation, virtual reality to video games. In this paper a survey of different 3D object reconstruction methods are presented. The methods are: Structure From Motion (SFM), shape-from-silhouette, voxel-coloring, evidence grids and variational methods.

Keywords: *Voxel coloring, Structure From Motion, silhouette.*

1 Introduction

Extensive research has been dedicated to solve the three-dimensional (3D) reconstruction problem. The objective is to obtain a three-dimensional representation of the observed scene that complies with certain constraints. Many approaches have been proposed in literature. Most of the approaches are based on the relative motion between views, silhouette (outline) of the object or color information on the surface of the object. This paper is organized as follows: Section 2.1 presents the Structure From Motion (SFM) approach. Section 2.2 introduces the shape-from-silhouette approach, Section 2.3 discusses voxel coloring method, Section 2.4 presents Evidence grids algorithm, Section 2.5 introduces variational methods and finally Section 3 is the conclusion.

2 Reconstruction Approaches

2.1 Structure From Motion

The structure from motion approach (SFM) refers to the estimation of the relative camera pose and the

3D reconstruction of a scene from at least two images taken from different view points. Since the seminal work of Longuet-Higgins [10] many SFM algorithms have been introduced [23, 4, 6, 11]. The main idea of the SFM algorithms is to minimize an objective function which represent the epipolar geometry constraint and recover the motion between views. The Structure From Motion method starts by detecting some image features like corners [6] or edges [1] in the two images. Then the *correspondence problem* [12] is solved by matching features that correspond to the projection of the same scene structure. The matching is usually based on cross-correlating image intensities [25] between views. The set of matched image features are used to compute an initial estimate of the epipolar geometry between the views. The constraint induced on the views by the initial estimate of the epipolar geometry is used in a guided feature matching step to refine the set of matched features. The epipolar geometry is calculated from the refined set of matched features and finally the motion between the views and the camera matrices are estimated. The 3D structure of the corresponding features can be obtained by using the projection matrices and applying triangulation [22] from at least two views of the scene as shown in Figure 1.

This approach can be applied on scenes that contain distinct features such as corners, lines or edges. However, for scenes with featureless objects other method like shape from silhouette are more appropriate.

2.2 Shape from silhouette

The shape from silhouette approach is applied efficiently on smooth textureless objects. The silhouette (outline) is the projection of points on the surface of the object at which the visual ray from the camera center is orthogonal to the surface normal. In general, the silhouettes of different views of a smooth object represent the projection of different curves in the ob-



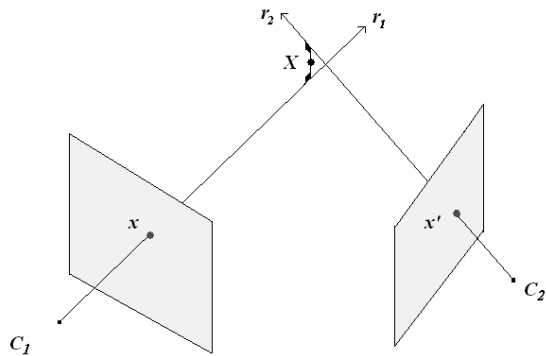


Figure 1: Reconstruction by Triangulation: two corresponding image points \mathbf{x} and \mathbf{x}' define two rays $r_1 = C_1\mathbf{x}$ and $r_2 = C_2\mathbf{x}'$. The equation of the rays r_1 and r_2 are known and the intersection can be computed to recover the 3D world point \mathbf{X} , in practice the rays will not intersect, thus the 3D point \mathbf{X} can be estimated as the minimum distance between the two rays.

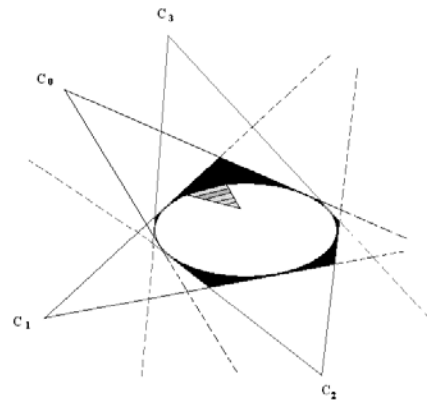


Figure 2: Reconstruction using silhouettes: The reconstruction of the 2D elliptic object using the four silhouettes is indicated by the dark area. Adding sufficient images, the elliptic object will be reconstructed. However, the concavity (hatched area) cannot be reconstructed regardless of the number of images used.

ject's 3D space. Thus silhouettes are view dependent in contrast to features, like corners [6], that are view independent. View dependence leads to the development of algorithms that do not assume rigidity of the scene unlike the SFM algorithms.

With the *shape-from-silhouette* strategy [2, 26], each view is segmented into silhouette and background points. The union of all visual rays emanating from all views and intersecting the image silhouette defines a generalized cone within which the 3D object must lie. If a sufficient number of views is used, the *visual hull* of the observed 3D structure is obtained [9]. However, as indicated by Laurentini [9], concavities in the object cannot be removed and the *visual hull* is an approximation of the true 3D structure of the reconstructed object. Figure 2 shows the reconstruction of a 2D elliptic object using four silhouettes. The generalized cones associated with the four images result in reconstruction represented by the dark area. By adding a sufficient number of images, the elliptic object will be reconstructed, however the concavity (the hatched area) cannot be reconstructed regardless of the number of images used. The visual hull of a 2D scene equals the convex hull; however, for 3D scenes the visual hull is included in the convex hull.

The silhouette image is represented as a binary image, with pixels classified as silhouette points or background points. Figure 3(a) shows some image silhouettes, Figure 3(b) shows the conic volume. To model the 3D structure of objects in the scene, the scene is first discretized into voxels in 3D space and then the 3D structure of the scene is modeled by the voxel occupancy description corresponding to the intersection of the conic volumes. The voxels are modeled using octree representation [3, 16, 21]. The 3D scene is initially represented by a coarse volume of voxels

and then each voxel is projected into all the input images and then tested if it belongs to the silhouettes. The voxel is processed based on three cases. First if the voxel belongs to all silhouettes, then the voxel is marked as opaque (i.e belongs to the object visual hull). If it does not intersect any silhouette, then the voxel is marked as transparent (i.e does not belong to the object visual hull). Finally if the voxel belongs to at least one silhouette, the voxel is subdivided into octants and the algorithm is applied recursively on the resulting sub-voxels.

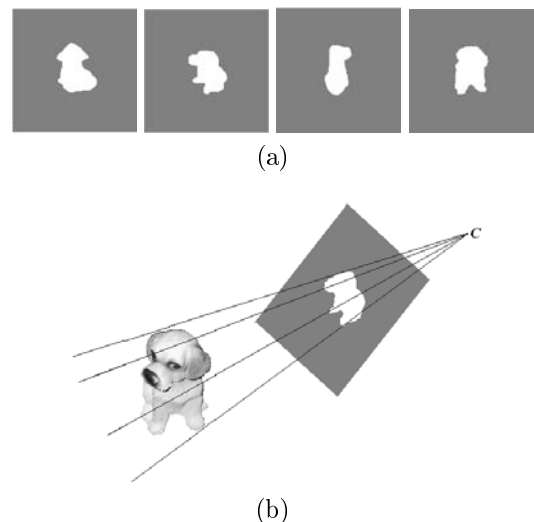


Figure 3: Shape from Silhouettes: (a) Image silhouettes (b) the conic volume.



2.3 Voxel coloring

While the shape-from-silhouette methods use only the binary images of the outline of the objects, the voxel-coloring methods make use of the image pixel colors. The photometric information captured by the set of input images can be used to recover the 3D structure of the scene. The *photo-consistency* [18] can be used to distinguish points on a surface of an object from other points on the scene. A point of a scene is said to be photo-consistent if the image pixel intensity, from each image in which this point is visible, equals the image irradiance of that point. Figure 4 shows two 3D points p_1 and p_2 visible from two cameras C_1 and C_2 . Point p_1 is located on the surface of an object, its projections on the two cameras at a_1 and a_2 have consistent pixel colors, thus point p_1 is said to be photo-consistent with the two images. Point p_2 lies above the ground plane, its projections on the two cameras at b_1 and b_2 have inconsistent pixel colors and p_2 is not photo-consistent.

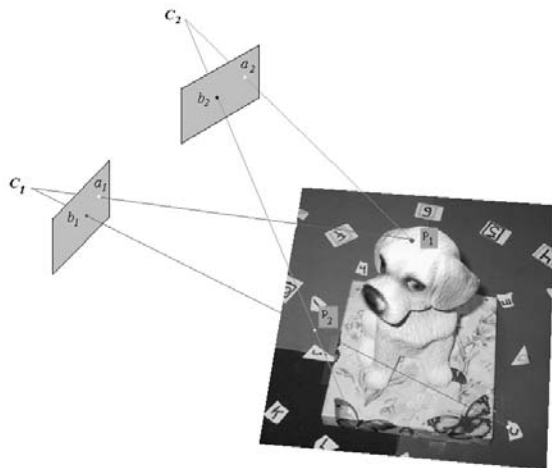


Figure 4: Color consistency can be used to distinguish points on a scene. Two 3D points p_1 and p_2 are visible from two cameras C_1 and C_2 . Point p_1 is located on the surface of the object, the two cameras see consistent colors on image points a_1 and a_2 . Point p_2 is located above the ground and the two cameras see inconsistent colors at points b_1 and b_2 .

The photometric information contained in each view can be used in the 3D reconstruction process by using the *voxel-coloring* strategy introduced by Seitz and Dyer [18]. The algorithm begins by dividing the scene space into small volumetric elements (the voxels) as shown in Figure 5. The scene is then visited and each initially opaque voxel is projected onto the input images and examined to determine if it belongs to an object surface. This is done by measuring the *photo-consistency* [18] constraint between the set of input images. The voxels that are found to be inconsistent are made transparent. The algorithm terminates when all the remaining opaque voxels are photo-

consistent. These voxels form a model that describes the 3D structure of the scene. The visibility of each voxel must be determined before performing color consistency. For a given voxel, the transparency of all the voxels that occlude it must be determined first. If a given voxel is not occluded by an opaque voxel, then this given voxel is considered to be visible. Seitz and Dyer [18] introduced the *ordinal visibility* constraint on the camera locations to simplify the voxel visibility computation. The voxels are visited in a near to far order relative to all the cameras. This can be achieved by placing the cameras on one side of the scene and then visit the voxels in planes at increasing distances from the cameras. For arbitrary camera placement, the voxels are visited in planes at increasing distances from the convex hull of the cameras.

Prock and Dyer [17] introduced a method to enhance the performance of the voxel-coloring approach [18]. The approach is based on a coarse-to-fine voxels strategy. Coarse-resolution voxels are used to represent the interior of large objects and the free space areas in the scene and fine voxels are used to represent objects surfaces. First, the scene is represented using coarse voxels and the voxel-coloring approach is applied. The resulting opaque voxels are then subdivided into eight smaller ones and another pass of voxel-coloring is executed. The algorithm is repeated until a certain defined resolution is achieved.

Kutulakos and Seitz [8] introduced an approach called *space carving* to reconstruct scenes using arbitrary camera placement. The approach is similar to the voxel-coloring approach, planes of voxels are scanned and evaluated for color consistency. The *ordinal visibility* of the scanned plane of voxels is maintained by sweeping a plane through the cameras and the images that are passed by the plane are used in the computation. To ensure photo-consistency of voxels with all input views, multiple sweeps are needed. Six passes on each voxel is used by sweeping a plane through the positive and negative directions of the 3D voxel coordinate system. As the algorithm runs, the voxels that are found to be inconsistent are made transparent (i.e are *carved*). When the algorithm terminates, the remaining set of opaque voxels form a model of the scene which is a superset of any other photo consistent model. This model is called the *photo hull*.

Both the SFM and the voxel-coloring methods require the projection matrices for all cameras to be known. However, for the voxel-coloring approach solving the correspondence problem for each 3D voxel is not required; instead each voxel is projected on all the images and the photo-consistency is calculated. On the other hand, solving the correspondence problem for the SFM is inevitable and it is challenging especially for smooth objects with constant radiance.



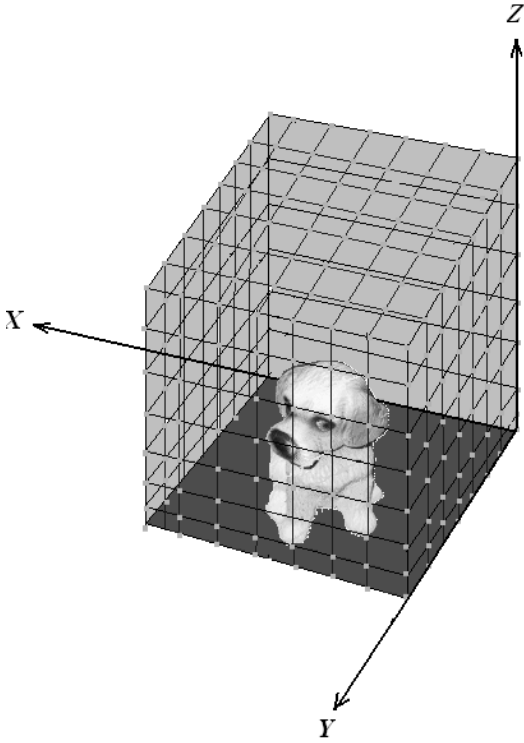


Figure 5: Voxel coloring: The scene is divided into voxels.

2.4 Evidence grids

The evidence grids approach, proposed by Moravec [14, 13], is used to represent regions of robot's surrounding using stereoscopic images. It integrates concepts from both voxel coloring and Structure From Motion (SFM) methods.

The 3D space is modeled as a three-dimensional array of cells (voxels) called the *evidence grid*. Each grid cell accumulates evidence about its occupancy. The cells are initialized to zero to indicate no evidence about the occupancy. After the reconstruction process, the grid cells are updated and the positive cells indicate occupied regions and the negative cells indicate free space. The approach [14] starts by collecting images using a calibrated stereo camera. First, features are detected in the two images and then correlated to obtain a set of corresponding features [12]. The 3D structure of each corresponding feature is calculated using triangulation [22] as discussed in the SFM approach (see Section 2.1). The cells that lie along the rays joining the calculated 3D point and the two cameras are assigned negative values (i.e free space) and a positive value (i.e occupied) on the 3D point. The final 3D structure of the scene is described by the occupied cells in the grid. To obtain a photo realistic model of the environment similar to the model obtained using the voxel coloring approach (see Section 2.3), the occupied cells are colored by back-projecting the original images.

2.5 Variational methods

Traditional 3D reconstruction algorithms for the stereo problem is to solve the *correspondence* problem first, then solve the *reconstruction* problem for each corresponding pair of image points using triangulation as discussed in Section 2.1. In their pioneer work, Faugeras and Keriven [5, 15] introduced a variational approach without separating the two problems. The method is based on 3D reconstruction in the context of a multi-view stereo problem, i.e. when several cameras of known perspective projection matrices are observing the same objects. Their work proposed functionals for matching to control the evolution of the surface. The functionals depends on cross-correlation of image intensities across pairs of views. To demonstrate the idea of the algorithm, consider the point \mathbf{M} on Figure 6. This point belongs to a surface S to be reconstructed and is projected onto two images \mathbf{I}_1 and \mathbf{I}_2 on points \mathbf{m}_1 and \mathbf{m}_2 respectively. The object S is represented in the neighborhood of point \mathbf{M} by its tangent plane. This tangent plane induces a homography transformation \mathbf{H} between points \mathbf{m}_1 and \mathbf{m}_2 . Thus, a rectangular window centered at point \mathbf{m}_1 in image \mathbf{I}_1 is mapped to a window, not necessarily rectangular, centered at \mathbf{m}_2 in image \mathbf{I}_2 . The cross-correlation between the two windows can be used to measure the similarity between image points \mathbf{m}_1 and \mathbf{m}_2 .

Notice that the homography \mathbf{H} is a function of the normal \mathbf{n} and the point \mathbf{M} . Based on this, an error criterion can be stated as follows:

Given a point \mathbf{S} on S with normal vector \mathbf{n} and projected at \mathbf{m}_1 in \mathbf{I}_1 , the correlation between images \mathbf{I}_1 and \mathbf{I}_2 at point \mathbf{m}_1 is defined to be:

$$\phi(\mathbf{S}, \mathbf{n}, \mathbf{m}_1) = \frac{\langle \mathbf{I}_1, \mathbf{I}_2 \rangle}{|\mathbf{I}_1| \cdot |\mathbf{I}_2|} \quad (1)$$

where $\langle \mathbf{I}_1, \mathbf{I}_2 \rangle$ denotes correlation between \mathbf{I}_1 and \mathbf{I}_2 . An error criterion for any point on the surface S can be written based on Equation (1) as follows:

$$C(\mathbf{S}, \mathbf{n}) = \int_S \phi(\mathbf{S}, \mathbf{n}) d\sigma \quad (2)$$

The integration over S is carried out with respect to the area element $d\sigma$.

Finally, the Euler-Lagrange equation of the variational problem in Equation (2) is written and the problem is numerically solved using the level sets methods [20].

Yezzi and Soatto [24] introduced a 3D reconstruction algorithm from a collection of images where the scene under observation is made of smooth surfaces with constant radiance. The algorithm starts with a rough estimate of the relative camera pose then the surface shape and motion parameters are simultaneously estimated. The method is based on a cost functional that minimizes the discrepancy between the projection of the 3D model onto the images and the



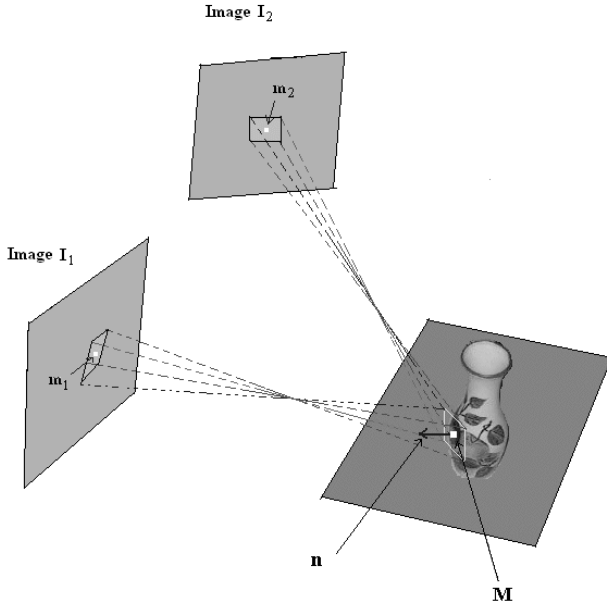


Figure 6: The square window on the first image is mapped to a window, not necessarily square, in the second image using the homography induced by the tangent plane to the surface S at point M .

measured projections captured by the collection of images.

The scene background and the object surface are assumed to have different radiance functions \mathbf{h} and \mathbf{f} respectively. This is an important assumption for the segmentation process during the surface reconstruction and pose estimation. The cost functional of the algorithm is a function of radiance functions \mathbf{f} and \mathbf{h} , surface S and camera pose g . For constant radiance functions \mathbf{f} and \mathbf{h} the cost function can be described as follows:

$$C(\mathbf{f}, \mathbf{h}, S, g) = C_{data}(\mathbf{f}, \mathbf{h}, S, g) + C_{geom}(S) \quad (3)$$

where C_{data} describes the discrepancy between measured images and the images predicted by the model and C_{geom} describes the smoothness of the surface. The Euler-Lagrange equations are derived for Equation (3) and then implemented using level set methods [20].

Sekkati and Mitiche [19] introduced a variational approach for recovering relative depth and 3D motion from a temporal sequence of images. The relation between the velocity of a point $\mathbf{P} = [X, Y, Z]^T$ in the 3D Euclidean space and the velocity of its image $\mathbf{p} = [x, y, f]^T$ in the image plane can be described by the basic equation of motion as follows [7, 22]:

$$\begin{cases} v_x = \frac{t_z x - t_x f}{Z} - w_y f + w_z y + \frac{w_x x y}{f} - \frac{w_y x^2}{f} \\ v_y = \frac{t_z y - t_y f}{Z} + w_x f - w_z x - \frac{w_y x y}{f} + \frac{w_x y^2}{f} \end{cases} \quad (4)$$

where $\mathbf{t} = [t_x, t_y, t_z]^T$ is the translational component

of motion and $w = [w_x, w_y, w_z]^T$ is the angular velocity. Under the assumption that the brightness from a point on a surface in the scene does not change during motion, the following gradient equation is satisfied:

$$E_t + \dot{x}E_x + \dot{y}E_y = 0 \quad (5)$$

where E_x , E_y and E_t are the spatio-temporal derivatives of the image brightness E . A rigid body motion constraint equation can be obtained based on Equations (4) and (5) as follows:

$$E_t + \mathbf{s} \cdot \frac{\mathbf{t}}{Z} + \mathbf{q} \cdot w = 0 \quad (6)$$

where \mathbf{s} and \mathbf{q} are both functions of x , y , f , E_x and E_y . The 3D motion of the image sequence can be recovered by minimizing the following cost functional:

$$C(\mathbf{t}, w, Z) = C_{motion} + C_{reg} \quad (7)$$

where C_{motion} represents the conformity to the rigid body motion in Equation(6) and C_{reg} is a regularization term to preserve the boundaries of the 3D interpretation.

Finally, as in [5, 15] and [24], the Euler-Lagrange equations for Equation (7) are derived and implemented using level set methods [20].

3 Conclusion

In this paper a survey of different 3D reconstruction algorithm is presented. The SFM [23, 4, 6, 11] approach performs efficiently on textured objects; however reconstructing objects with constant radiance remains a challenge for this approach.

The shape-from-silhouette method [2, 26] performs efficiently on scenes with textureless objects. Images in this approach need to be segmented into foreground and background regions, for voxel coloring method [18] such segmentation is not required. On the other hand the voxel-coloring approach requires the projection matrix to be known; however solving the correspondence problem is not needed.

The voxel-based approaches[2, 26, 18, 14, 13, 5, 15] provide smoother 3D models compared to the SFM method. However the SFM method is computationally more efficient. The number of 3D grid points (voxels) has an important impact on the accuracy of the model as well as on the computational cost. More voxels results in more accurate model but computationally more expensive to reconstruct. Faster implementation of the voxel-based method remains a subject of active research.

References

- [1] J. Canny. A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):679–698, Nov 1986.



- [2] C.H. Chian and J.K. Aggarwal. Model reconstruction and shape recognition from occluded contours. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11:372–389, 1989.
- [3] C.H. Chien and J.K. Aggarwal. Volume / surface octrees for the representation of three-dimensional objects. *Computer Vision, Graphics, and Image Processing*, 36(1):100–113, 1986.
- [4] O. Faugeras. *Three-Dimensional Computer Vision, A geometric Viewpoint*. MIT Press, Cambridge, MA, 1996.
- [5] O. Faugeras and R. Keriven. Variational principles, surface evolution, PDE's, level set methods and the stereo problem. *IEEE Trans. Image Processing*, 7(3):336–344, 1998.
- [6] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey vision Conf*, pages 147–151, 1988.
- [7] B. Horn and B. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.
- [8] K. N. Kutulakos and S.M Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.
- [9] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(2):150–162, 1997.
- [10] H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from projections. *Nature*, 293:133–135, 1981.
- [11] M.I.A Lourakis and S.C. Orphanoudakis. Visual detection of obstacles assuming a locally planar ground. In *Proc. 3rd Asian Conf. on Computer Vision*, 2:527–534, 1998.
- [12] D. Marr and T. A. Poggio. Cooperative computation of stereo disparity. *Science*, 194(4262):283–287, October 1976.
- [13] M.C. Martin and H.P. Moravec. Robot evidence grids. Technical Report 06, Carnegie Mellon University, Robotics Institute, Pittsburgh, PA, USA, 96.
- [14] H.P. Moravec. Robot spatial perception by stereoscopic vision and 3d evidence grids. Technical Report 34, Carnegie Mellon University, Robotics Institute, Pittsburgh, PA, USA, 96.
- [15] S. Osher and N. Paragios. *Geometric Level Set Methods in Imaging, Vision, and Graphics*. Springer, 2003.
- [16] M. Potmesil. Generating octree models of 3d objects from their silhouettes in a sequence of images. *Computer Vision, Graphics, and Image Processing*, 40(1):1–29, 1987.
- [17] A.C. Prock and C.R. Dyer. Towards real-time voxel coloring. In *Proc. Image Understanding Workshop*, pages 315–321, 1998.
- [18] S.M. Seitz and C.R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *IEEE conf. on Computer Vision and Pattern Recognition*, pages 1067–1073, 1997.
- [19] H. Sekkati and A. Mitiche. Dense 3d reconstruction of image sequence: a variational approach using anisotropic diffusion. In *Proceedings of the 12th International Conference on Image Analysis and Processing, (ICIAP'03)*, 2003.
- [20] J.A. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, 1996.
- [21] R. Szeliski. Rapid octree construction from image sequences. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 58(1):23–32, 1993.
- [22] E. Trucco and A. Verri. *Introductory Techniques for 3D Computer Vision*. Prentice-Hall, New Jersey, 1998.
- [23] R. Y. Tsai, T.S. Huang, and W.L. Zhu. Estimating three-dimensional motion parameters of a rigid planar patch. *IEEE Acoustic Speech Signal Processing*, 30(4):525–534, Aug 1982.
- [24] A. Yezzi and S. Soatto. Structure from motion for scenes without features. In *IEEE conf. on Computer Vision and Pattern Recognition*, pages 554–559, 2003.
- [25] Z. Zhang. A new and efficient iterative approach to image matching. In *ICPR*, Jerusalem, Israel, 1994.
- [26] J.Y. Zheng. Acquiring 3d models from sequences of contours. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(2):163–178, 1994.

