Structure and Motion from Silhouettes

Kwan-Yee K. Wong and Roberto Cipolla
Department of Engineering, University of Cambridge,
Trumpington Street, Cambridge, CB2 1PZ, UK
[kykw2|cipolla]@eng.cam.ac.uk

Abstract

This paper addresses the problem of recovering structure and motion from silhouettes. Silhouettes are projections of contour generators which are viewpoint dependent, and hence do not readily provide point correspondences for exploitation in motion estimation. Previous works have exploited correspondences induced by epipolar tangencies, and a successful solution has been developed in the special case of circular motion (turntable sequences). However, the main drawbacks are (1) new views cannot be added easily at a later time, and (2) part of the structure will always remain invisible under circular motion. In this paper, we overcome the above problems by incorporating arbitrary general views and estimating the camera poses using silhouettes alone. We present a complete and practical system which produces high quality 3D models from 2D uncalibrated silhouettes. The 3D models thus obtained can be refined incrementally by adding new arbitrary views and estimating their poses. Experimental results on various objects are presented, demonstrating the quality of the reconstructions.

1. Introduction

Silhouettes (alternatively referred to as outlines, profiles or apparent contours) are often a dominant image feature, and can be extracted relatively easily and reliably. They provide rich information about both the shape and motion of an object, and are indeed the only information available in the case of smooth textureless surfaces. Nonetheless, structure and motion from silhouettes has always been a challenging problem [11, 3, 2, 17, 1, 5, 10, 15]. Unlike corners, silhouettes are projections of contour generators [3] which are viewpoint dependent, and hence do not readily provide point correspondences. A traditional approach to the problem is to search for *epipolar tangencies* [16, 2]. However, such a search involves a nonlinear optimization with nontrivial initialization, and requires the presence of 7 or more epipolar tangencies. Recently, a practical solution has been found in the special case of circular motion [15], which requires only 2 epipolar tangencies. However, the main drawbacks are (1) new views cannot be added easily at a later time, and (2) part of the structure will always remain invisible under circular motion.

In this paper, we present a complete and practical system for generating high quality 3D models using 2D silhouettes from uncalibrated views taken under circular and arbitrary general motion. Only the 2 outer epipolar tangents, which are always present in a pair of views of an object, are required for the motion estimation. The incorporation of arbitrary general views reveals information which is concealed under circular motion, and overcomes the drawbacks of using circular motion alone.

Section 2 presents the theoretical background for structure and motion from silhouettes. The algorithms and implementations are described in Section 3. Section 4 shows the experimental results, and conclusions are given in Section 5.

2. Theoretical Background

Silhouettes are projections of contour generators, and do not readily provide point correspondences. A *frontier point* [9, 2, 4] is the intersection of two contour generators and lies on an epipolar plane which is tangent to the surface (see fig. 1). It follows that a frontier point will project onto a point in the silhouette which is also on an epipolar tangent [16].

Epipolar tangencies thus provide point correspondences which can be exploited for motion estimation. Theoretically, if 7 or more epipolar tangencies are available, the epipolar geometry between 2 views can be estimated, and the camera intrinsic parameters can then be used to recover the relative motion [12, 6]. However, when the epipolar geometry is not known, the localization of the epipolar tangencies involves a nonlinear optimization with nontrivial initialization. The unrealistic demand for a large number of epipolar tangencies and the presence of local minima also

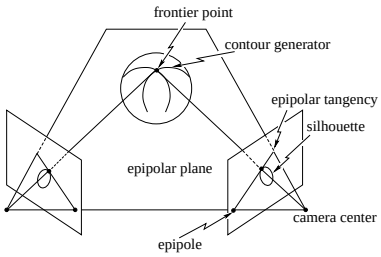

Figure 1. A *frontier point* is the intersection of two contour generators and lies on an epipolar plane which is tangent to the surface. It follows that a frontier point will project onto a point in the silhouette which is also on an epipolar tangent.

make this approach impractical.

In this paper, a simple method which uses only the 2 *outer epipolar tangents* for motion estimation is introduced. The outer epipolar tangents correspond to the 2 *epipolar tangent planes* which touch the object, and are always available in any pair of views, except when the baseline passes through the object (see fig. 2). The use of the outer epipolar tangents, which are guaranteed to be in correspondence, avoids degenerate cases due to self-occlusion and greatly simplifies the matching problem (see fig. 3).

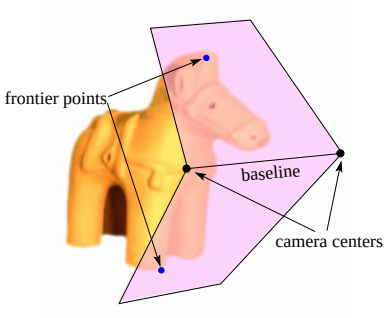

Figure 2. The outer epipolar tangents correspond to the 2 epipolar tangent planes which touch the object, and are always available in any pair of views, except when the baseline passes through the object.

Circular motion can be used to determine a few viewpoints reliably by having a good initialization that avoids local minima, which is the key to a successful solution. In the case of circular motion, the 3 main image features, namely the image of the rotation axis $\mathbf{l}_\mathrm{s}$, the horizon $\mathbf{l}_\mathrm{h}$ and a special vanishing point $\mathbf{v}_x$ (see [15] for details), are fixed throughout the sequence and satisfy

$$\mathbf{v}_x \cdot \mathbf{l}_\mathrm{h} = 0, \quad (1)$$

$$\mathbf{v}_x = \mathbf{K}\mathbf{K}^\mathrm{T}\mathbf{l}_\mathrm{s}, \quad (2)$$

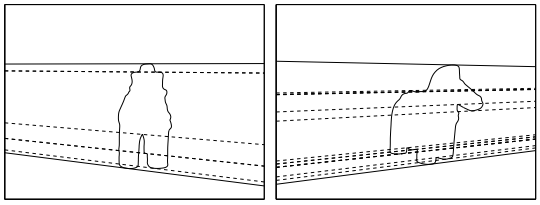

Figure 3. Two discrete views showing 17 epipolar tangents, of which only 6 are in correspondence. The use of the 2 outer epipolar tangents (in solid lines), which are guaranteed to be in correspondence, avoids degenerate cases due to self-occlusion, and greatly simplifies the matching problem.

where $\mathbf{K}$ is the 3×3 camera calibration matrix. The fundamental matrix can be parameterized explicitly in terms of these features [19, 7, 15], and is given by

$$\mathbf{F} = [\mathbf{v}_x]_\times + \kappa \tan\frac{\theta}{2}(\mathbf{l}_\mathrm{s}\mathbf{l}_\mathrm{h}^\mathrm{T} + \mathbf{l}_\mathrm{h}\mathbf{l}_\mathrm{s}^\mathrm{T}), \quad (3)$$

where θ is the angle of rotation, and κ is a constant which can be determined from the camera intrinsic parameters. θ is the only parameter which depends on the particular pair of views being considered. When the camera intrinsic parameters are known, 2 parameters are enough to fix $\mathbf{l}_\mathrm{s}$ and $\mathbf{v}_x$. Since $\mathbf{v}_x$ must lie on $\mathbf{l}_\mathrm{h}$, only 1 further parameter is needed to fix $\mathbf{l}_\mathrm{h}$. As a result, a sequence of N images taken under circular motion can be described by $N + 2$ motion parameters (2 for $\mathbf{l}_\mathrm{s}$, 1 for $\mathbf{l}_\mathrm{h}$ and the $N - 1$ angles). By exploiting the 2 outer epipolar tangents, the N images will provide $2N$ (or 2 when $N = 2$) independent constraints on these parameters, and a solution will be possible when $N \geq 3$.

In this paper we show how circular motion will generate a *web of contour generators* around the object (see fig. 4), which can be used for registering any new arbitrary general view. Given an arbitrary general view, the associated contour generator will intersect with this web and form frontier points. If the camera intrinsic parameters are known, the 6 motion parameters (3 for rotation and 3 for translation) of the new view can be fixed if there are 6 or more frontier points on the associated contour generator. This corresponds to having a minimum of 3 views under circular motion, each providing 2 outer epipolar tangents to the silhouette in the new general view (see fig. 5).

3. Algorithms and Implementations

In this paper, the cubic B-spline snake [3] is chosen for the automatic extraction of silhouettes from the image sequence. Cubic B-spline snakes provide a compact representation of silhouettes of various complexity, and can achieve

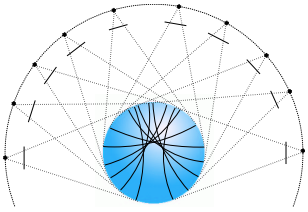

Figure 4. The circular motion will generate a *web of contour generators* around the object, which can be used for registering any new arbitrary general view.

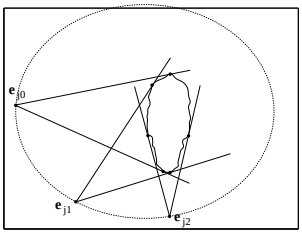

Figure 5. Three views from circular motion provide 6 outer epipolar tangents to the silhouette in the new general view for estimating its pose.

sub-pixel localization accuracy. Its parameterization also facilitates the localization of epipolar tangents.

The motion estimation proceeds as an optimization which minimizes the reprojection errors of epipolar tangents. For view i and view j, a fundamental matrix $\mathbf{F}_{ij}$ is formed from the current estimate of the motion parameters, and the epipoles $\mathbf{e}_{ij}$ and $\mathbf{e}_{ji}$ are obtained from the right and left nullspaces of $\mathbf{F}_{ij}$. The outer epipolar tangent points $\mathbf{t}_{ij0}$, $\mathbf{t}_{ij1}$ and $\mathbf{t}_{ji0}$, $\mathbf{t}_{ji1}$ are located in view i and j respectively (see fig. 6). The reprojection errors are then given by the geometric distances between the epipolar tangent points and their epipolar lines [14],

$$d_{ijk} = \frac{\mathbf{t}_{jik}^{\mathrm{T}}\mathbf{F}_{ij}\mathbf{t}_{ijk}}{\sqrt{(\mathbf{F}_{ij}\mathbf{t}_{ijk})_1^2 + (\mathbf{F}_{ij}\mathbf{t}_{ijk})_2^2}}, \quad (4)$$

$$d_{jik} = \frac{\mathbf{t}_{jik}^{\mathrm{T}}\mathbf{F}_{ij}\mathbf{t}_{ijk}}{\sqrt{(\mathbf{F}_{ij}^{\mathrm{T}}\mathbf{t}_{jik})_1^2 + (\mathbf{F}_{ij}^{\mathrm{T}}\mathbf{t}_{jik})_2^2}}. \quad (5)$$

For a sequence of N images taken under circular motion, the image of the rotation axis and the horizon are initialized approximately, and the angles are arbitrarily initialized. The cost function is given by

$$C_{cm}(\mathbf{x}) = \frac{1}{4}\sum_{i=1}^{N}\sum_{j=i+1}^{min(j+3,N)}\sum_{k=1}^{2} d_{ijk}(\mathbf{x})^2 + d_{jik}(\mathbf{x})^2, \quad (6)$$

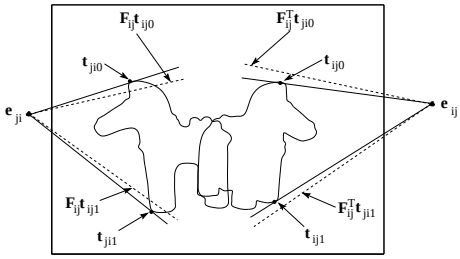

Figure 6. The motion parameters can be estimated by minimizing the reprojection errors of epipolar tangents, which are given by the geometric distances between the epipolar tangent points and their epipolar lines.

where $\mathbf{x}$ consists of the $N+2$ motion parameters.

For arbitrary general motion, the 6 motion parameters can be initialized by roughly aligning the projection of the 3D model built from the estimated circular motion with the image. The cost function of general motion for view j is given by

$$C_j(\mathbf{x}') = \sqrt{\frac{\sum_{i=1}^{N} f_{ij}\sum_{k=1}^{2} d_{ijk}(\mathbf{x}')^2 + d_{jik}(\mathbf{x}')^2}{4\sum_{i=1}^{N} f_{ij}}}, \quad (7)$$

where $\mathbf{x}'$ consists of the 6 motion parameters. f_{ij} is 0 if the baseline formed with view i passes through the object, otherwise it is 1.

After motion estimation, a volumetric model of the object can be constructed by octree carving algorithm [18] using the silhouettes. Surface triangles can then be generated from the octree using the marching cubes algorithm [13] and texture for each triangle is taken from the view which gives the largest projected area.

Instead of using texture-mapping which sometimes produces color inconsistency between neighboring triangles, a model with smooth color transition can be obtained by computing the RGB values for each vertex in the mesh. The normal vector of each vertex is taken as the mean of the normal vectors of the neighboring triangles. The color is then computed as the weighted average of the color values from all views. The weighting factor is given by

$$w_i = \begin{cases} -\mathbf{n}\cdot\mathbf{v}_i & (\text{if visible}) \\ 0 & (\text{otherwise}) \end{cases}, \quad (8)$$

where $\mathbf{n}$ is the normal vector of the vertex and $\mathbf{v}_i$ is the viewing direction of view i.

The complete process of generating 3D model from 2D silhouettes is summarized in algorithm. 1.

Algorithm 1 Generation of 3D model from silhouettes

```
extract the silhouettes using cubic B-spline snakes;

initialize l_s, l_h and the N − 1 angles for the circular motion;
while not converged do
   compute the cost for circular motion using (6);
   update the N + 2 motion parameters to minimize the cost;
end while

for each arbitrary general view do
   initialize the 6 motion parameters;
   while not converged do
      compute the cost for the general motion using (7);
      update the 6 motion parameters to minimize the cost;
   end while
end for

carve the octree using the silhouettes;
form surface triangles by marching cubes algorithm;
generate texture-maps;

for each vertex in the mesh do
   compute normal vector;
   compute color as weighted average of color values from views where
   the vertex is visible;
end for
```

4. Experimental Results

The first experimental sequence consists of 15 uncalibrated images of a Haniwa (large hollow baked clay sculpture placed on ancient Japanese burial mounds), of which the first 11 images are taken under circular motion of the Haniwa (see fig. 7). The 3D model built from the circular motion alone is shown in fig. 10. The gaps between the legs are not carved away since they never appear as part of the silhouettes, and textures are missing in areas (top and bottom) which are invisible under circular motion. Figure 11 shows the refined model, with great improvements in both shape and textures, after incorporating 4 arbitrary general views. Figure 12 shows the same refined model with colored vertices instead of texture-mapping.

The second experimental sequence consists of 13 uncalibrated images of a human head, of which the first 10 images are taken under circular motion of the camera (see fig. 8). The 3D model built from the circular motion alone is shown in fig. 13, with textures missing at the top of the head and under the chin. The refined model after incorporating 3 arbitrary general views is shown in fig. 14 with texture-mapping and in fig. 15 with colored vertices respectively. The specularities from sun light and the automatic exposure adjustment of the camera caused color inconsistency in the image sequence, which makes the texture-mapped 3D model look inferior.

The third experimental sequence consists of 9 images of a Haniwa (see fig. 9) in front of a calibration grid, and is used for quantitative evaluation. Each view in the sequence

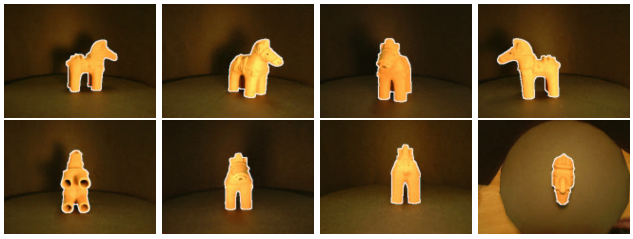

Figure 7. Top: four images from an uncalibrated sequence of a Haniwa under circular motion. Bottom: four additional images of the Haniwa under general motion, featuring the bottom, front, rear and top views.

Figure 8. Eight images from an uncalibrated sequence of a human head.

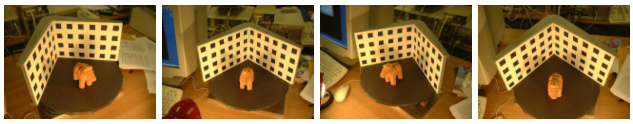

Figure 9. Four images from the calibrated sequence of a Haniwa used for quantitative evaluation of the general motion estimation algorithm.

was calibrated independently using the 3D data from the grid. The algorithm for estimating the pose of an arbitrary general view, as described in Section. 3, was then used to register the 9^{th} view using different subsets of the first 8 images. The results are presented in table 1 which shows the reprojection errors of the corner features from the calibration grid. Though the errors for the motion estimated using epipolar tangents are not as small as those from the calibration grid, the results are indeed very good since only 6–16 epipolar tangent points have been used, compared with 192 corners used in the case of the calibration grid.

5. Conclusions and Future Work

In this paper, we have presented a complete and practical system for generating high quality 3D models from 2D silhouettes. The input to the system is a sparse, uncalibrated

Figure 10. 3D model of the Haniwa built from the circular motion sequence. The gaps between the legs are not carved away since they never appear as part of the silhouettes, and textures are missing in areas (top and bottom) which are invisible under circular motion.

Figure 11. 3D model of the Haniwa after incorporating the 4 additional views, showing great improvements in both shape and textures.

Figure 12. Same model as in fig. 11, but with colored vertices instead of texture mapping.

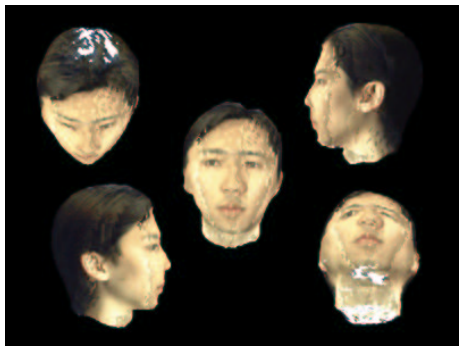

Figure 13. 3D model of the human head built from the circular motion sequence. Textures are missing at the top of the head and under the chin.

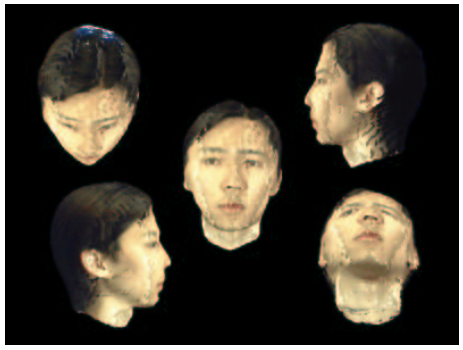

Figure 14. 3D model of the human head after incorporating the 3 additional views. The whole model is now covered with textures. The color inconsistency is due to the specularities from sun light and the automatic exposure adjustment of the camera.

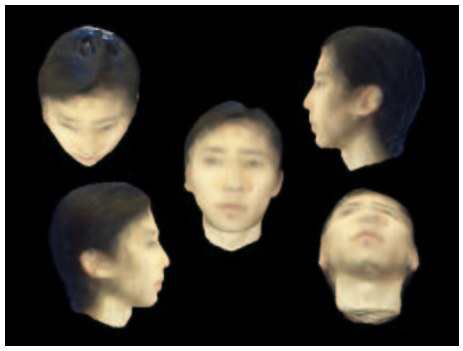

Figure 15. Same model as in fig. 14, but with colored vertices instead of texture mapping.

Table 1. Reprojection errors (in pixels) of the 192 corner features from the calibration grid.

motion estimated from: silhouettes in views	no. of tangent pts	reprojection error: rms (in pixels)
1–3	6	1.1824
1–4	8	0.9407
1–5	10	0.8858
1–6	12	0.7311
1–7	14	0.7856
1–8	16	0.7963
ground truth	corners	rms (in pixels)
calib. grid	192	0.3853

image sequence of an object under both circular and general motion. The incorporation of arbitrary general views reveals information which is concealed under circular motion, and greatly improves both the shape and textures of the 3D models. It also allows incremental refinement of the 3D models by adding new views at any time, without the need of setting up the exact, identical scene carefully. The registration of general motion using circular motion (or alternatively 3 or more known views) avoids the problem of local minima which exists in every general motion estimation using silhouettes. Since only silhouettes have been used in both the motion estimation and octree carving, no corner detection nor matching is necessary. This means that the system is capable of reconstructing any kind of objects, including *smooth* and *textureless* surfaces. Experiments on various objects have produced convincing 3D models, demonstrating the practicality of the system.

We are currently extending the system to handle image sequence of *approximate* circular motion, like hand-held video sequence around a sculpture. The motion parameters obtained by assuming a circular motion may be used to initialize a full optimization for the true general motion. We would also like to extend the system to recover surface reflectance [8, 20], so as to produce photo-realistic 3D models under different lighting conditions.

References

[1] K. Åström and F. Kahl. Motion estimation in image sequences using the deformation of apparent contours. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 21(2):114–127, Feb 1999.

[2] R. Cipolla, K. Åström, and P. Giblin. Motion from the frontier of curved surfaces. In *Proc. 5th Int. Conf. on Computer Vision*, pages 269–275, Boston, USA, Jun 1995.

[3] R. Cipolla and A. Blake. Surface shape from the deformation of apparent contours. *Int. Journal of Computer Vision*, 9(2):83–112, 1992.

[4] R. Cipolla and P. J. Giblin. *Visual Motion of Curves and Surfaces*. Cambridge University Press, Cambridge, 1999.

[5] G. Cross, A. W. Fitzgibbon, and A. Zisserman. Parallax geometry of smooth surfaces in multiple views. In *Proc. 7th Int. Conf. on Computer Vision*, pages 323–329, Kerkyra, Greece, Sep 1999.

[6] O. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig. In G. Sandini, editor, *Proc. 2nd European Conf. on Computer Vision*, pages 563–578, Santa Margherita Ligure, Italy, May 1992. Springer–Verlag.

[7] A. W. Fitzgibbon, G. Cross, and A. Zisserman. Automatic 3D model construction for turn-table sequences. In *3D Structure from Multiple Images of Large-Scale Environments, European Workshop SMILE'98*, pages 155–170, Freiburg, Germany, June 1998. Springer–Verlag.

[8] P. Fua and Y. G. Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *Int. Journal of Computer Vision*, 16:35–56, Sep 1995.

[9] P. J. Giblin, F. E. Pollick, and J. E. Rycroft. Recovery of an unknown axis of rotation from the profiles of a rotating surface. *J. Opt. Soc. America A*, 11(7):1976–1984, Jul 1994.

[10] T. Joshi, N. Ahuja, and J. Ponce. Structure and motion estimation from dynamic silhouettes under perspective projection. *Int. Journal of Computer Vision*, 31(1):31–50, 1999.

[11] J. J. Koenderink. What does the occluding contour tell us about solid shape? *Perception*, 13:321–330, 1984.

[12] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[13] W. E. Lorensen and H. E. Cline. Marching cubes: a high resolution 3D surface construction algorithm. In M. C. Stone, editor, *Proc. SIGGRAPH 87, Computer Graphics Proceedings, Annual Conference Series*, pages 163–170, Anaheim, CA, Jul 1987.

[14] Q.-T. Luong and O. Faugeras. The fundamental matrix: Theory, algorithm, and stability analysis. *Int. Journal of Computer Vision*, 17(1):43–75, 1996.

[15] P. R. S. Mendonça, K.-Y. K. Wong, and R. Cipolla. Camera pose estimation and reconstruction from image profiles under circular motion. In D. Vernon, editor, *Proc. 6th European Conf. on Computer Vision*, volume II, pages 864–877, Dublin, Ireland, Jun 2000. Springer–Verlag.

[16] J. Porrill and S. B. Pollard. Curve matching and stereo calibration. *Image and Vision Computing*, 9(1):45–50, 1991.

[17] S. Sullivan and J. Ponce. Automatic model construction and pose estimation from photographs using triangular splines. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 20(10):1091–1096, Oct 1998.

[18] R. Szeliski. Rapid octree construction from image sequences. *CVGIP: Image Understanding*, 58(1):23–32, Jul 1993.

[19] T. Vieville and D. Lingrand. Using singular displacements for uncalibrated monocular visual systems. In B. Buxton and R. Cipolla, editors, *Proc. 4th European Conf. on Computer Vision*, volume II, pages 207–216, Cambridge, UK, April 1996. Springer–Verlag.

[20] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In A. Rockwood, editor, *Proc. SIGGRAPH 99, Computer Graphics Proceedings, Annual Conference Series*, pages 215–224, Los Angeles, California, Aug 1999. Addison Wesley Longman.