SHAPE FROM DARKNESS: Deriving Surface Information from Dynamic Shadows

John R. Kender*
Earl M. Smith

Department of Computer Science
Columbia University
New York, New York 10027

1 Abstract

We present a new method, *shape from darkness*, for extracting surface shape information based on object self-shadowing under moving light sources. It is motivated by the problem of human perception of fractal textures under perspective. One-dimensional dynamic shadows are analyzed in the continuous case, and their behavior is categorized into three exhaustive shadow classes. The continuous problem is shown to be solved by the integration of ordinary differential equations, using information captured in a new image representation called the *suntrace*. The discretization of the one-dimensional problem introduces uncertainty in the discrete *suntrace*; however it is successfully recast as the satisfaction of $8n$ constraint equations in $2n$ unknowns. A form of relaxation appears to quickly converge these constraints to accurate surface reconstructions; we give several examples on simulated images. The shape from darkness method has two advantages: it does not require a reflectance map, and it works on non-smooth surfaces. We conclude with a discussion on the method's accuracy and practicality, its relation to human perception, and its future extensions.

2 Introduction

We present a new, active method for obtaining shape information from low level cues. It exploits the information implicit in the shadows that an object or an object part casts upon itself or another object. In spirit, it is most like the photometric stereo method of Woodham (Woodham, 1981), in that it requires control over illuminant position. However, it also extends the existing work on shadow geometry of Shafer (Shafer, 1985) and others, and gives additional insight into the nature of shadows, especially in the cases where the objects are neither polyhedra nor smooth, or where the shadows are dynamically changing. The method has two major advantages. It appears to work best for textured objects, that is, where existing methods fail most badly. And it is more robust than existing methods, in that it requires little a priori information about a surface's reflectance. Further, it illustrates the inherent utility--and complexity--of static or dynamic shadow-based cues for any integrated vision system, whether active or passive.

3 Historical Background

The method, which can be called *shape from darkness*, was motivated by an interest in the human perception of fractal textures. As Pentland has shown (Pentland, 1984), the fractal dimension of textured surfaces is a powerful feature on which the segmentation of an image can be based. He further observed that the image of a single fractal surface viewed under perspective has non-constant fractal dimension. It is conjectured that this change in measured feature is closely related to the change in overall local surface orientation of the surface with respect to the observer. If this is the case, then fractal dimension can serve as a basis for a "shape from fractal" method, similar to other gradient-based shape from x methods.

However, the mathematics behind such relationships appear formidable. This is because the observed change in fractal dimension appears to be due to the increasing self-occlusion of the fractal surface as it is viewed at increasingly oblique angles. That is, unlike an airborne observer of a mountain range, an observer down in the foothills sees very little of the mountain peaks: he sees mostly the sides of foothills. The mathematical difficulty stems from the intractability of the threshold-like non-linear functions that express the nature of object occlusion; the difficulties are similar to the ones faced when trying to integrate object segmentation with standard shape from x methods.

*This research was supported in part by ARPA grant #N00039-84-C-0165, by a NSF Presidential Young Investigator Award, and by Faculty Development Awards from AT&T, Ford Motor Co., and Digital Equipment Corporation.

Nevertheless, the problem does have the following analogue, which ultimately suggested the method reported here. It is that self-occlusion is very similar to shadowing: were a light source moved to the observer's position, the self-occluded areas would now be the ones in shadow. Thus, instead of attempting to investigate the effect that varying surface orientations have on observed fractal properties (or, equivalently, the effect that varying observer positions have), one can explore the effects that varying light source positions have on the generation of a fractal's shadows. Ideally, one would like to look into the shadows in order to see what information has been lost.

Generating and analyzing shadow information allows for several computational efficiencies. Essentially, when working with shadows, one is doing rendering and shading under extreme conditions. The capture of shadow information from real imagery or the generation of shadows synthetically both result in binary imagery. Instead of collecting shading information that has a range of values, one obtains a characteristic function instead: zero means shadow, one means illuminated. Simple thresholding of actual imagery is usually all that is required, and the synthetic casting of shadows is a straightforward computation. The imagery that results can be seen as extreme shape from shading in another sense. A synthetic shadow image can be obtained in the standard graphic rendering way by first thresholding the reflectance map: all gradients which reflect any light at all are set to one, and the remainder of the map stays at zero (for self-occluding). What results when an image is rendered with such a map is an image with extreme contrast; indeed, the contrast cannot be more extreme.

Recovering the depth or orientation of those surface fragments that have been shadowed is clearly a difficult task given only one shadow image. As with many other problems in vision, many influences are conflated into the simple image observable, the shadow. The beginning of a shadow is determined not only by surface orientation and illuminant direction, but also by the absence of any prior surface to overshadow it. The termination of a shadow depends on the relative heights and orientations of both the shadowing and shadowed surface. Deconflating these influences in a single image is not necessarily impossible; it depends on the additional information and assumptions one also brings to the task. For example, if it is known that the surface is that of a hemisphere, its position and radius are easily recovered, even without knowledge of the illuminant direction. Less restrictive assumptions, such as the surface having a band-limited fourier spectrum (and therefore "smooth" in exactly this sense of smooth), may also admit to solutions, perhaps in a form analogous to the Logan theorem characterizing a signal by its zero-crossings (Logan, 1977). But still weaker assumptions, such as the surface simply being twice differentiable, probably do not lead to solutions at all. This is because smoothness as defined by differentiability is the assumption implicit in true shape from shading, and true shape from shading depends heavily on the amount of curvature in the reflectance map (Lee, 1985); the thresholded reflectance map has none.

4 Problem Formalization

The shape from darkness problem is more straightforward to solve by using multiple images. The observer and the objects can be held stationary, obviating any image-to-image correspondence problem, and what is moved is the light source, in a manner similar to photometric stereo. Photometric stereo usually can be done with three illuminant positions, although four is the usual number used in practice in order to prevent exactly the problem discussed here: objects in self-shadow. It is apparent that even four shadow images is woefully inadequate for shape from darkness under reasonable surface assumptions. Thus, the problem is relaxed to allow a fixed number of illuminant positions, the exact count and location of which are to be determined. The added complexity of increased imagery is mitigated in part by its binary nature, and in part by the lack of any necessity to calibrate the shadow reflectance map, since the latter is determined solely by the

illuminant direction. One only needs to define the location of the shadow terminator orientations.

For simplicity in the discussion that follows, the problem is further reduced to its natural one-dimensional subproblem. That is, the algorithms presented here will discuss the recovery of a planar curve rather than a surface, given illuminants that lie in the plane of the curve. (The extension of the method to the full two dimensional case, including a discussion of the degrees of freedom of illuminant placement, is sketched later.) Thus, we assume that depth is a function solely of x, z=f(x), rather than z=f(x,y), and that the illuminants lie wholly within the xz plane. Note that photometric stereo has a similar one-dimensional analogue, with one-dimensional reflectance maps that are functions of curve derivative rather than of surface gradient. In one-dimensional photometric stereo, three lights are necessary to prevent objects--here, curves--from self-shadowing.

5 The Continuous Problem

It is instructive to consider the shape from darkness problem as a continuous problem first. Assume that the illuminant is an infinitely distant point source, and that the observer is infinitely far in the positive z direction. (Thus, instead of investigating the surface properties of a fractal seen under perspective, we are now exploring the recovery of curve information from shadows generated under parallel illumination.) Given that the illuminant will appear in many orientations, it will be convenient to identify the illuminant with the sun, the positive direction of the x axis with the east, illumination at zero slope with dawn, illumination at positive slopes with morning, and illumination from the positive z axis with noon; often these terms are more immediate and compact.

As shown in the figure, it is easy to show that under these conditions all curve points fall into one of three classes of dynamic shadow behavior under increasing morning illumination, with analogous classes in the afternoon. A point either can become illuminated because it gradually is moved out from self-shadowing, or it can be always illuminated, or it can become illuminated because it gradually moves out from a cast shadow. These definitions can be made precise, at the given points:

A *minus* point m has f'(m) >=0 such that for all x, x>m implies f(x) <= f(m) + f'(m)(x-m). (Implicitly, f''(m) <= 0.) Intuitively, a minus point can only be in shadow when it is (or would be) self-shadowed. It becomes illuminated precisely at the time of day when the rising illuminant's slope is equal to f'(m). When f(m) becomes illuminated, points to the immediate west of m remain in shadow; therefore, in the direction of illumination the transition at m is from illumination into darkness. This terminator travels west with increasing illumination, and crosses descending values of f. Note that the shadow is caused by light grazing f(m), and it is therefore diffuse, especially at low illuminant slopes. Such a point is therefore called minus for five negatively flavored reasons (a sixth becomes apparent shortly): its second derivative is negative, its terminator goes from light to dark, the terminator travels west, the terminator descends, and the shadow is not sharp.

A *zero* point z is such that for all x, x>z implies f(x) <= f(z). (Implicitly, f'(z) <= 0.) Intuitively, a zero point is never shadowed (in the morning), not even at dawn. It becomes illuminated when the rising illuminant has slope equal to zero. It never experiences a terminator: it is characterized by zero shadow and zero change.

A *plus* point p is every other point. Negating and manipulating quantifies yields: either f'(p) <= 0 and it is not a zero point, or f'(p) >= 0 and it is not a minus point. Intuitively, a plus point can only be shadowed due to cast shadows. It becomes illuminated when the rising illuminant grazes a minus point at m (illuminant slope is f'(m)), such that f(m) = f(p) + f'(m)(m-p). When f(p) becomes illuminated, points to the immediate east of p remain in shadow; therefore, in the direction of illumination the transition at p is from darkness into illumination (thus, plus). This terminator travels east (plus) with increasing illumination. Note that the shadow is caused by occlusion, and is therefore sharp (plus). (However, f''(p) is not necessarily positive, and the terminator does not necessarily cross ascending values of f.)

The function f can therefore be partitioned into segments and the segments labeled by their shadow class. The grammar of segment labels is simple; in the morning it is given by the regular expression $((+-)^*0)^*$. Such strings have three significant transitions. Plus to minus occurs at f' = 0 with f' at a local maximum. Minus to zero occurs at f' = 0 with f at a local maximum. Minus to plus occurs at curious "second grazing" points, those points m where f'(m) is equal to the illuminant slope, but where there is also a p>m with f'(p) also equal to the illuminant slope, and f(p) = f(m) + f'(m)(p-m). (The fourth transition, zero to plus, appears to have no special significance.)

5.1 The Continuous Suntrace

Quantitative reconstruction can be based on the integration of the derivative information intrinsic in the minus points. The reconstruction requires an additional representation of image information, called the *suntrace*, from which the requisite derivative information is obtained.

The suntrace is a mapping from the domain of the original curve into (morning) illumination slopes. For each x, it records the slope at which the value f(x) first became illuminated. The suntrace is a function of x, since a given f(x) can become illuminated only once. Depending on the underlying curve, the suntrace may be unbounded: although the entire curve must be illuminated no later than noon, noon corresponds to an unbounded illumination slope.

Since zero points are illuminated at dawn, they have suntrace values identically zero; see Figure 1. Minus points are likewise easy to detect and label: they are exactly those points (in the morning) with negative (minus) suntrace derivatives, since their terminators move west with increasing illuminant slope. What remains are the plus points; they have positive (plus) suntrace derivatives.

5.2 Solution Using ODEs

Given a morning suntrace, the underlying curve can be partially reconstructed. A contiguous curve segment with minus labels can be integrated into a function segment by using the suntrace value of the point as the value of f' at the point. The segment, however, must "float" at an unknown height until it is given an absolute height by the appropriate constant of integration.

By definition, the function values of all plus points can be determined relative to the position of their corresponding minus points that shadow them. For a plus point p, the calculation is based on the relation f(m) = f(p) + f'(m)(m-p), where the corresponding minus point m is found in the suntrace as the least m greater than p that has the same illumination slope, f'(m), that p has. Entire contiguous segments of plus edges can therefore be fixed in space, and joined to their integrated minus segment

The now completed plus-minus complexes can themselves be joined one to another at their common "second grazing points" (that is, at minus-plus transitions). In this way, long, self-consistent segments of the curve result, but with each "floating" with respect to a constant of integration; see Figure 2.

The fuller recovery can never be made since a simple morning suntrace provides no information about zero points. Their relative and actual depths can attain arbitrarily high values, and any self-consistent segments separated by zero points can freely float relative to each other, as long as the slope of the intervening zero segments remain negative.

Pinning down the constant of integration and restricting the behavior of zero points can be achieved by using a second suntrace, usually the afternoon suntrace which maps illumination slopes from noon to dusk. It is apparent that the only point that can be labeled a zero point for both suntraces is the global maximum. All other points are shadowed at least once and can therefore be assigned a function value relative to some constant. What results, within the accuracy of the suntrace and the integration, is a reconstruction of the underlying curve with depth values relative to a single constant of integration: the global maximum.

6 The Discrete Problem

Discretizing the shape from darkness problem requires some care. The heart of the difficulty is twofold. A discrete suntrace, however fine, can only give upper and lower bounds to function derivatives for minus points. Further, given digitization, it is not always clear where the shadowing function values in the discrete suntrace really ought to be. Indeed, as the method below describes, occasionally two different function values will serve to set the bounds on the derivatives, one for the upper bound and one for the lower bound, since the true minus point may be somewhere between them.

Rather than attempt to approximate a derivative for each minus edge, the following method attempts to maintain solution accuracy by calculating the exact upper and lower bounds the solution could have, given the discretization of the suntrace that it starts with.

6.1 The Discrete Suntrace

Any given function value f(x) will become illuminated only once. Thus, there will be a time of morning, t, at and before which f(x) is in shadow, and after which, at t+1, it becomes illuminated. Call the function value that shadowed f(x) at time t (but could not shadow it at time t+1) the *last shadower*. The last shadower may not be the only function value that fails at time t+1 to shadow f(x); call the function value responsible for the (potential) shadow the *failing shadower*. It is not hard to see that for illumination in the morning, the west-to-east order of the function values is f(x), failing shadower, and last shadower, although there may be many other undistinguished function values scattered amongst these three. These two shadowers generate important constraints on the upper and lower bounds of f(x); more interestingly, f(x) itself feeds back constraints on the upper and lower bounds of its shadowers, too. It is important to note that these two shadowers of f(x) may be the last and failing shadowers of other function values, too; their roles at these other points may even be reversed. See Figure 3.

The shape from darkness method begins by collecting from the discrete suntrace, for every element x in the domain of the curve, information about such shadowers. The last shadower of f(x) is found in the morning in the following way. If f(x) first became illuminated at time t+1, the last shadower of f(x) was the nearest eastern illuminated neighbor to f(x) at time t. The failing shadower of f(x) is the nearest eastern illuminated neighbor at time t+1.

Fortunately, such information can be collected in one pass through the suntrace. Assuming both a morning and afternoon suntrace, each element x of the domain will gather eight pieces of information: for each of the four morning or afternoon last or failing shadowers, it stores their position and the time of their shadowing (t or t+1).

6.2 The Eight Constraints per Point

Given this information, each point in the domain affects and is affected by these four critical shadowers. Each point therefore participates in eight constraints, four to do the affecting, and four to be affected by. Given that the morning and afternoon suntraces are completely symmetrical, there are only four basic conceptual relations: forward or backward constraints on upper or lower bounds. The forward constraints propagate constraint information in the direction of the illuminant; the backward constraints propagate it against the illuminant.

The forward constraints are based on the following observations. At point x, x's upper bound can be no higher than the projected shadow of the upper bound of its last shadower. (If x's upper bound were any higher, x would not be shadowed at time t). Similarly, at point x, x's lower bound can be no lower than the projected shadow of the lower bound of its failing shadower. (If x's lower bound were any lower, x would instead be shadowed at time t+1).

In the morning, the forward constraint equations are therefore:

UPPER $u(x) <= u(ls(x)) - (ls(x)-x)*sls(x)$

LOWER $l(x) >= l(fs(x)) - (fs(x)-x)*sfs(x)$

where u(.) and l(.) represent the upper and lower limits in effect at any time, ls(.) and fs(.) are the coordinates of the last shadower and failing shadower, and sls(.) and sfs(.) are the illumination slopes at the times of last shadow and failing shadow.

The backward constraints are a bit trickier, but it is their feedback that seems to account for the method's power. Consider the upper bound at x. Since the failing shadower must fail to shadow x, the upper bound of the failing shadower is limited by the height at which it just barely fails to shadow x; the maximum allowable height for the failing shadower occurs when x itself is at its maximum. (If the failing shadower's upper bound were higher, it would instead shadow x.) This height can be determined by backprojecting the upper bound of x along the slope in effect at the failing shadow time, t+1. Similarly, consider the lower bound at x. Since the last shadower must successfully shadow x, the lower bound of the last shadower is limited by the depth at which it just barely succeeds in shadowing x; the minimum allowable depth for the last shadower occurs when x itself is at its minimum. (If the last shadower's lower bound were smaller, it would instead fail to shadow x.) This height can be determined by backprojecting the lower bound of x along the slope in effect at the last shadow time, t. See Figure 4.

In the morning, the backward constraint equations are therefore:

UPPER $u(fs(x)) <= u(x) + (fs(x)-x)*sfs(x)$

LOWER $l(ls(x)) >= l(x) + (ls(x)-x)*sls(x)$

Four similar constraints apply to the information gathered for x from the afternoon suntrace.

It is surprising that these appear to be all the constraints possible (aside from the trivial constraint that $u(x) > l(x)$). Other relationships between the upper and lower bounds of x, upper and lower bounds of its last shadower, and upper and lower bounds of its failing shadower, do not appear to be constraining. For example, if x's upper bound decreases, it has no effect on the upper bound of its last shadower.

6.3 Solution Using Relaxation

The specific family of constraints that result from a given suntrace have a complex interrelated structure. It is not apparent whether there is any special solution method applicable to this problem in general, or even for well-defined subclasses of curves. There are $8n$ inequalities in $2n$ unknowns, and there is a well-defined objective function to minimize: that is, the sum, over all x, of $u(x) - l(x)$).

Although the problem might be solved using linear programming, a more attractive solution method is the use of a version of relaxation. Conceptually this consists of a number of successive iterations, in each of which the eight constraint equations are successively applied to each point x in the domain. If the application of any constraints results in better estimates for u(x) or l(x), they are updated. As in the continuous case, the only valid initial values are those of the global maximum (the only point labelled zero in both suntraces); its upper and lower limits are set arbitrarily to a pleasant value (say, zero) before the relaxation begins.

In practice, convergence seems very rapid. Unlike some relaxation algorithms, updating is based on thresholds, so upper and lower bounds are only altered if they are moved closer together. The method is therefore more likely to terminate when it recognizes a lack of measurable progress.

7 Experimental Results

In the experiments that follow, some of the generalities of the algorithm were made particular. For ease of comparing the final reconstructed curve to the original, the global maximum of the reconstruction was initialized to its true known height. Sun positions were simulated at constant slope increment; thus, sun angles in the morning linearly increase in tangent. (Under this scenario, the sun literally rises, rather than travels an arc!) This policy of constant increment seems to be closely related to the encouraging accuracy obtained in the final processing step, where the final estimate of the curve is defined to be the curve midway between the computed upper and lower bounds.

Each of these series of test images shows the following.

The first figure of a series is the original curve, with its morning and evening suntraces. The domain of the original curve is aligned with the domain of the suntraces. Both suntraces have the axes for increasing sun slope pointing toward the curve. Thus, on all suntraces, the line nearest the curve is pure black, indicating all pixels have been illuminated.

The second figure of a series is a record of the constraint processing. Initial estimates for upper and lower bounds as propagated from the global maximum have gradually approached each other, subject to the suntrace data.

The third figure of the series shows final upper and lower bounds, the original curve, and the superimposed best estimate.

The first series is an image of a self-similar mountain. It is approximately 300 points wide by 85 points peak-to-peak. The suntrace was taken at increments of .1, that is, at approximately four degrees, to a maximum of 30 increments. The final estimate has a cumulative total error of less than 68 (about .2 error per pixel, average), and a maximum single point error of less than 1.2.

The second series is the same image, but with a suntrace increment of 1: that is, the first non-dawn suntrace is taken at 45 degrees, and only four increments are possible. Although not a realistic test, it demonstrates more visibly the method and its results, especially the goodness of the final estimate even under extremely severe conditions.

The third series demonstrates the applicability of the processing to very smooth imagery: a semicircle of radius 50, again under 30 increments of .1 each. Maximum error occurs at the extreme left and right of the "table", although reconstruction error within the circle is no more than 0.5.

8 Discussion

It appears that the accuracy of the final estimate is surprisingly good, and may be related to the use of constant illumination slope increment. Choosing the midway curve is guaranteed to minimize worst case error, since the midpoint can never be off more than half the available range.

8.1 Performance

Aside from the empirical data given above, little is known about the theoretic performance of the algorithms except in two worst cases. In terms of accuracy, the worst case image occurs in a monotonically decreasing function with positive curvature (as in $z = 1/(x+c)$). Here, points at the extreme asymptotic end have little opportunity for feedback, so the range between upper and lower bounds is virtually the same as the initial forward constraints, length*(slope(t+1)-slope(t)); if slopes increase in constant increments, this is simply length*increment. In terms of convergence, it appears that certain square wave trains takes n iterations, where n is the number of pulses in the train.

Shape from darkness has several advantages, most notably that it can exploit the surface information implicit in a class of dynamic shadows, with very little restrictions placed on the class of surfaces being shadowed: they need not be smooth. In particular, it can probably be useful in increasing the accuracy with which finely textured surfaces are viewed, especially under oblique illumination. It can also exploit smart cameras that run-length encode the incoming binary shadow imagery, but the exact information content of a shadow image, especially with respect to the information content in a gray scale image, remains to be explored.

8.2 Practicality

The utility of the method depends upon the extent to which shadowed imagery can be accurately obtained. This does not necessarily imply a completely controllable artificial light source: natural sources such as the sun can be used if there is concurrent accurate slope (or time of day) information. In effect, the method establishes an upper bound on the error of reconstruction for *any* series of shadowed imagery, artificial or otherwise. Like other shape from x methods, it is best seen as one of many possible sources of surface information.

8.3 Relation to Human Perception

The complexity of the data interaction does suggest why humans do not appear to derive much surface information from dynamic shadows. The necessity to store, in effect, an entire suntrace is probably excessive. On the other hand, if our earth rotated much faster (say, once every three seconds), there may have been more reason for natural systems to develop at least an approximate solution to the shape from darkness problem.

8.4 Extensions

The method admits of many extensions. The application of the method to real imagery must address the difficulties of specularity, mutual illumination, and diffuse shadows. However, in a robot environment, much of the environment can be structured to make the problem easier. For example, having the knowledge that the object is on a fixed table at a given depth can aid in the setting of lower bounds.

The extension to two-dimensional surfaces is probably the most critical. The problem can probably be decomposed for parallel processing in ways beyond the trivial one of partitioning the images in strips parallel to the illuminant direction; it may even be done in a hierarchical way. Selecting optimal sun positions with two degrees of freedom is challenging, but may reduce to two simple perpendicular transits. The problem is especially acute if sun and observers are allowed to be near, and observers are allowed to view off the normal axis; this is once again the original problem of fractals under perspective.

9 Acknowledgements

Paul Douglas provided an early version of the code for the generation of suntrace information, and coined the name of the representation.

References

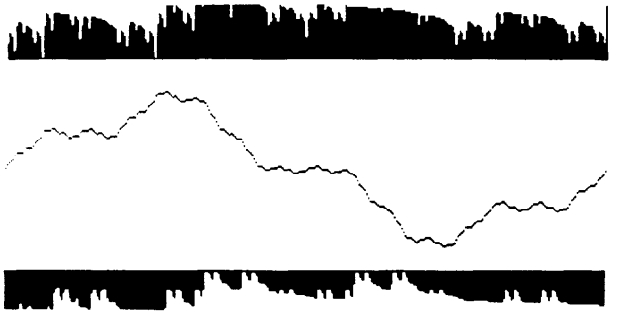
[1] Lee, D.
A Provably Convergent Algorithm for Shape from Shading.
In *Proceedings of the DARPA Image Understanding Workshop*, pages 489-496. December, 1985.

[2] Logan, B.F.
Information in the Zero-Crossings of Bandpass Signals.
Bell System Technical Journal 56:487-510, 1977.

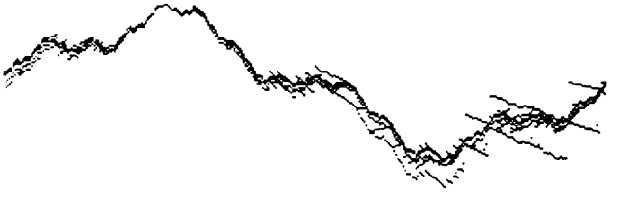
[3] Pentland, A.P.
Fractal-Based Description of Natural Scenes.
IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6(6):661-674, November, 1984.

[4] Shafer, S.A.
Shadows and Silhouettes in Computer Vision.
Kluwer Academic Publishers, Hingham, MA, 1985.

[5] Woodham, R.J.
Analysing Images of Curved Surfaces.
Artificial Intelligence 17(1-3):117-140, August, 1981.



First series: Original curve and suntrace



First series: Constraint propagation

First series: Final bounds and estimate

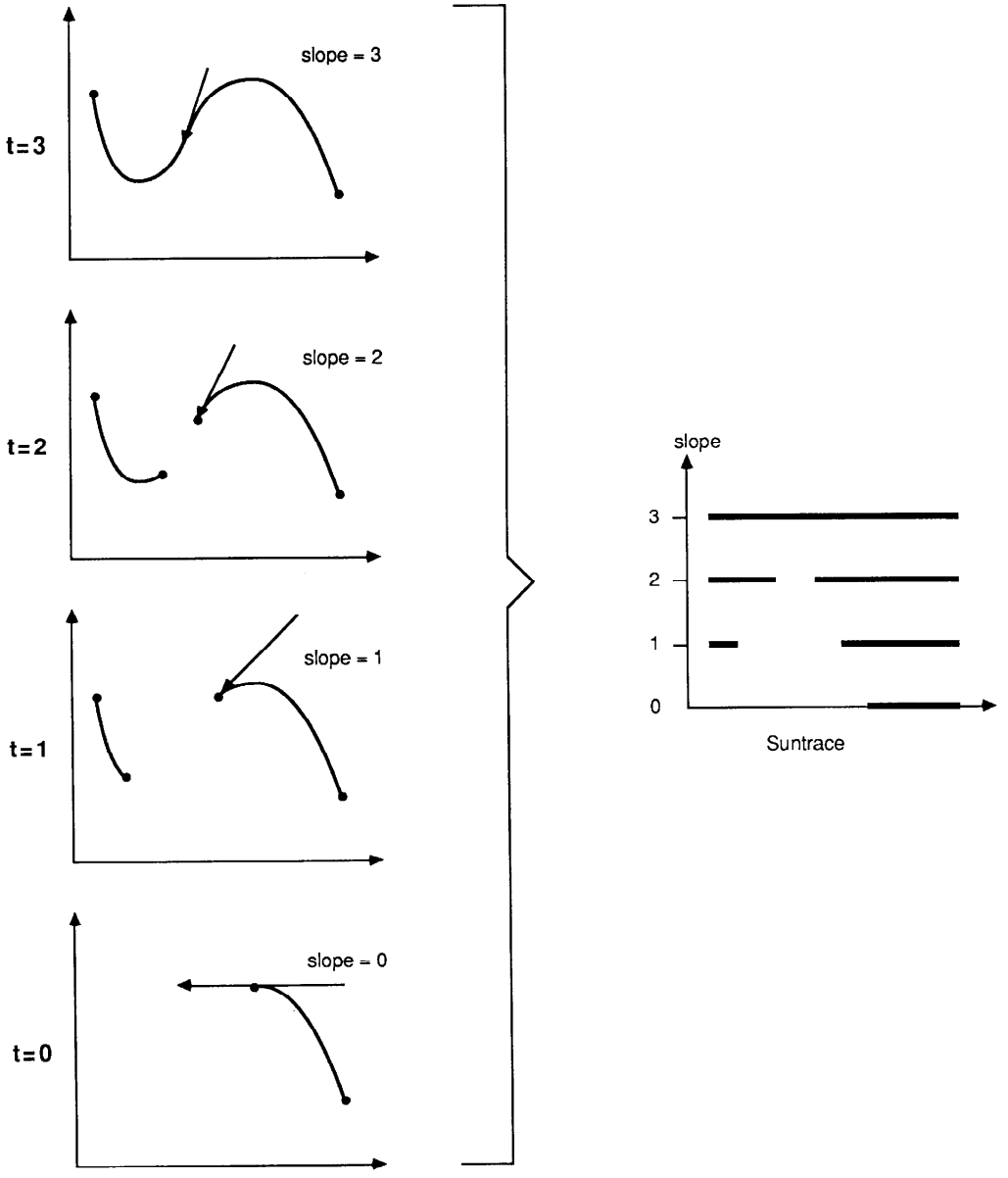


Figure 1

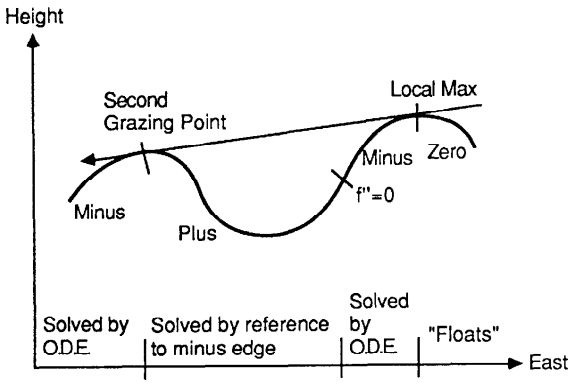


Figure 2

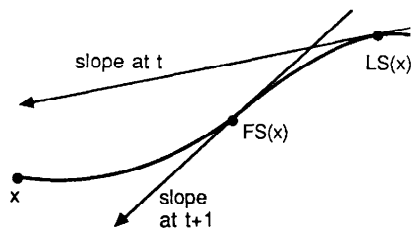


Figure 3

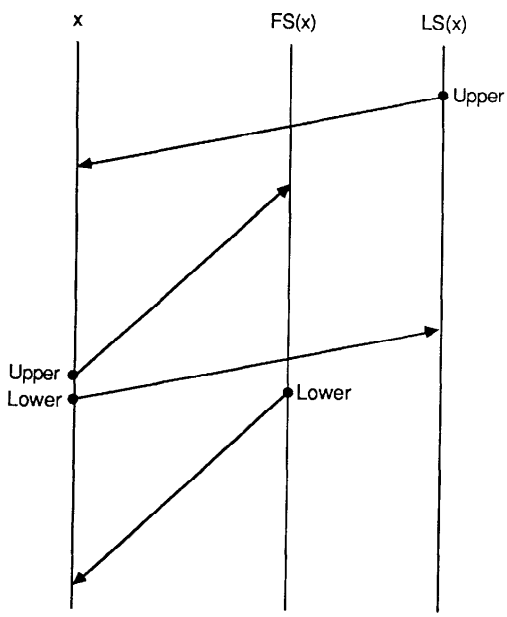
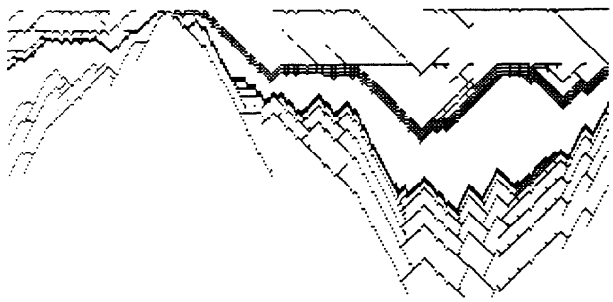
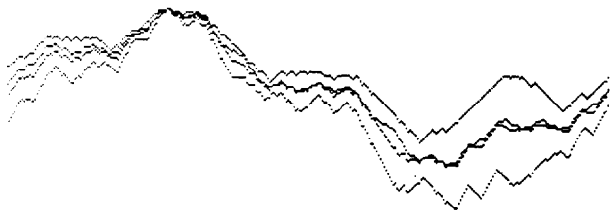


Figure 4

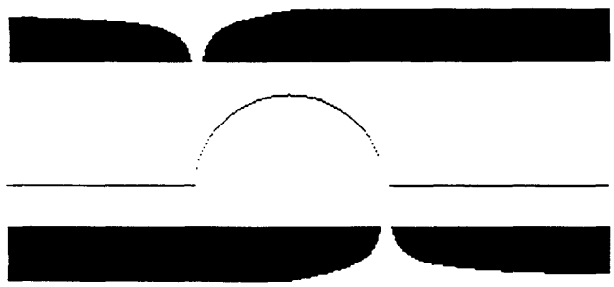
Second series: Original curve and suntrace



Second series: Constraint propagation



Second series: Final bounds and estimate



Third series: Original curve and suntrace

Third series: Constraint propagation

Third series: Final bounds and estimate