# Accurate and Consistent Reconstruction of Illumination Functions Using Structured Sampling

George Drettakis and Eugene Fiume
{dret|elf}@dgp.toronto.edu
Department of Computer Science, University of Toronto, Toronto, CANADA M5S1A4

## Abstract

*The study of common classes of diffuse emitters, such as planar convex polygons, reveals several interesting properties of the functions of illumination these emitters cast on receiver surfaces. Some properties, such as the position of the maximum and the curvature are of particular interest for sampling and reconstruction of illumination across receivers. A computationally efficient approach is presented that identifies these properties, and uses them to select samples of illumination. In addition these properties are used to determine upper bounds on the error due to linear and quadratic interpolants. These bounds are then used to adaptively subdivide the non-uniform sampling grid, resulting in accurate reconstruction. Results show that the method reduces the error compared to uniform approaches, and produces more consistent animated sequences.*

## 1. Motivation

In every global illumination algorithm, it is necessary to represent radiance as it varies across a surface in the environment that receives light. Most approaches achieve this with a grid of elements on every surface. Radiance is usually collected at the centre or at the vertices of these elements. The subdivision of these grids is typically a user-defined parameter that specifies minimum element size [9]. Some methods adaptively subdivide the grid based on the variation of illumination at these predefined sample points, while most depend ultimately on user intervention as a necessary final step.

These approaches have produced impressive results but nonetheless often result in element grids that are much finer than necessary in some places and overly coarse in others. The result is wasted computation while calculating the radiance values at each element and there is no guarantee of the quality or the error level incurred in such approaches, since the initial grid subdivision can cause significant undersampling errors. Finally, in the cases of animation, errors can appear as objects change relative positions, since previous approaches may give inconsistent results when the geometry changes.

The study of the behaviour of certain emitter classes such as convex polygons reveals several important functional properties. The algorithm presented in this paper takes advantage of this structure of illumination functions in a common class of emitters. A non-uniform, adaptive sampling strategy is developed that is then used to create a piecewise polynomial representation of radiance over a receiver surface. The goals of the approach are *efficiency*, achieved by decreasing the number of samples and therefore illumination function evaluations; *accuracy*, since the algorithm is based on bounding the error in specific regions; and finally *consistency* during animation, which is achieved by tracking certain important characteristics of the illumination function.

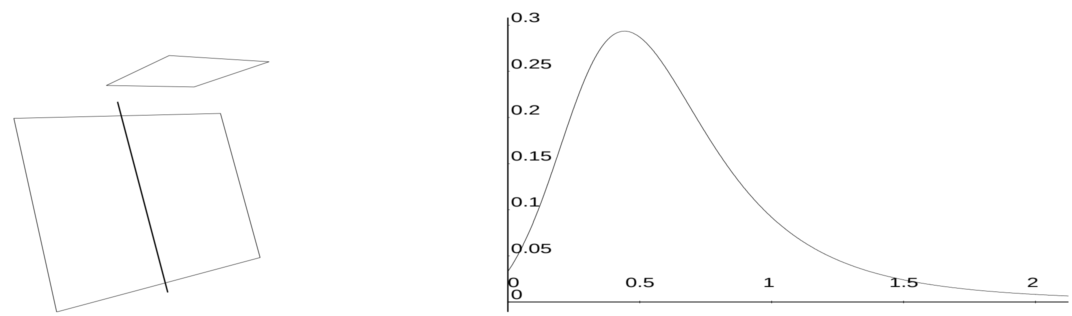
The method can be used in traditional radiosity-based algorithms as an alternative to the patch/element representation of radiance. It facilitates better illumination function quality, as well as relieving the user of the burden of manual patch size adjustment.

## 2. Previous Work

Previous algorithms have mainly used the "radiosity gradient" approach to adaptively subdivide the element grid [3]. In this approach the illumination values are examined at neighbouring elements on the grid, and if the difference between the values is larger than some predefined threshold, the elements are subdivided. More recent work refines the grid based on predetermined geometrical considerations [1], power transfer [5], or view dependent criteria [13]. These approaches are to a large extent more concerned with the light transfer calculations, and less with the display of radiance.

Other approaches use piecewise polynomial representations on triangular grids [9][12], but their main goal is to deal with the problems introduced by shadow boundaries. Similar considerations due to shadows are presented in [6][7]. Higher order interpolants are considered, but the emphasis is on the accuracy of light transfer calculations in the context of a finite element approach. Salesin et al. [12], perform cubic reconstruction given a mesh of samples. The use of cubic reconstruction seems more suited to the situations that include shadow boundaries, since we believe that for the unoccluded case presented here, linear and quadratic interpolants are sufficient. However, this choice depends on the computational/quality trade-offs of a specific application.

None of the above approaches take the characteristics of illumination functions into account, even though in most of them only the limited class of convex polygonal emitters is being considered. Campbell and Fussell [2], identify the existence of a single maximum in unoccluded regions from simple light sources. This information is used to guide sampling by subdividing the regions between maximum and minimum values. No justification of the existence of the maximum was given however. In addition, the overall function behaviour was not examined, and no subsequent effort was made to achieve a good fit to the function.




(a) (b)

**Figure 1. Rectangular Source and Corresponding Illumination on a Linear Subdomain**

## 3. Illumination Function Behaviour

By observation, it can be seen that illumination due to planar sources often has general characteristics that do not change from one source to another. In particular, simple convex polygonal sources seem to have one maximum, and are "radially" decreasing everywhere else. Take for example the configuration of a rectangular light source shown in Figure 1(a). We examine the illumination function defined on the vertical line down the middle of the receiver. Even though the receiver is oriented at an angle with respect to the emitting source, we see that the illumination increases smoothly towards the maximum and the decreases slowly away from it. This property is called *unimodality*, and we define it formally as follows.

**Definition**: *A function $f(x, y)$ is unimodal if and only if f has only one maximum, and the restriction of $f(x, y)$ to a linear subdomain $ax + by + c = 0$, also has only one maximum and is monotonically decreasing as a function of distance from that maximum.*

This property does not hold for all polygonal sources. If two polygonal unimodal sources are connected by a long thin strip, the combined function will have two separate maxima close to the maxima of the original sources. As will be shown in what follows, it is extremely important to identify the behaviour of illumination functions if possible. In the cases in which we can determine or assume unimodality, the use of attendant function properties allows us to construct consistent and accurate sampling strategies. Conversely, there exist some cases of non-convex sources that are unimodal, such as bowtie polygons that are only very slightly concave. Due to the symmetry and simplicity of geometry, some special cases of light sources can be shown to be unimodal. One such case is a disc source lighting an arbitrarily oriented planar receiver.

## 3.1 Proof of Unimodality for a Disc Light Source

Consider the configuration shown in Figure 2(a), where the disc shown is a uniform diffuse emitter. The point $P$ and the disc define a unique sphere. It has been shown that the illumination in the direction $OP$ from any point $P$ on the sphere (but below the disk) is equal, and depends only on the radius of the sphere [10]. We call any such sphere an *isolux* sphere. Define an isolux sphere $S$ with radius $R$. $H(R) = R + |OC|$ is the distance from the disc to the point directly opposing the centre (see Figure 2(a)) and is proportional to $R$. If the disc has radius $a$, and luminosity $L$, then the illumination[1] at any point on the sphere is given as follows:

$$E_s(R) = L\frac{a^2}{H(R)^2 + a^2} \tag{1}$$

To prove the unimodality property, we transform our environment so that the receiver plane is embedded in the plane $z = 0$. We define $E(P)$ to be the value of illumination at the point $P(x, y)$. Eq. (1) gives an expression for the illumination in direction $\overline{OP}$. To find the illumination with respect to the surface normal it is necessary to scale by the cosine of the angle $\theta$ formed between $OP$ and the normal to the surface $N_s$ (see Figure 2(b)). We will initially examine the function $E(x, y)$ and then demonstrate that the cosine scale factor preserves the decreasing nature of the function.

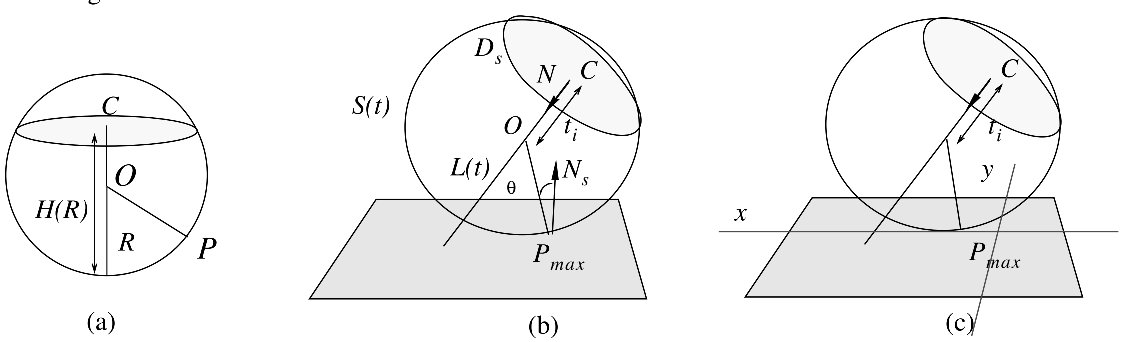



**Figure 2. Disc light source properties**

**Theorem**: *For a disc light source $D_s$ centred at $C = (c_x, c_y, c_z)$, with radius $a$, and outgoing unit normal to the surface $N = (n_x, n_y, n_z)$ and for the receiver plane $z = 0$, the illumination function $E(x, y)$ is unimodal.*

**Proof:** All the isolux spheres defined by the disc $D_s$ have centres that are positioned on the line $L(t) = (c_x, c_y, c_z) + t(n_x, n_y, n_z)$ [10] (see Figure 2(b)). Define $S(t_i)$ as the sphere centred at $O = L(t_i)$ for a specific $t = t_i$ (see Figure 2(b)). Call $R_s(t_i)$ the radius of the sphere.

**Lemma 1**: *There exists only one value $t_{min}$ of $t$ on the line $L(t)$ such that the corresponding unique sphere $S(t_{min})$ is tangential to the plane $z = 0$ at point $P_{max}$, thus having only one point in common with the plane.*

If the equation of the sphere is substituted into the plane $z = 0$, and if we require that the radius of the resulting circle is 0, we have two possible values of $t_{min}$. It can be shown from the geometry of the problem that only one of these roots is acceptable.

Any value $t_i$ less than $t_{min}$ corresponds to an isolux sphere $S(t_i)$ that does not intersect the receiving plane. The radius of the resulting isolux sphere is an increasing function of $t_i$. From Eq. (1), we know that the illumination value of the isolux sphere is an decreasing function of $R_s$, and thus a decreasing function of $t_i$. Consequently, of all the isolux spheres that have at least one point in common with the receiving plane, the sphere $S(t_{min})$ has the smallest radius, and consequently the highest value of illumination on the receiver plane. Therefore the function $E(x, y)$ has a single maximum at the point $P_{max}$ at which $S(t_{min})$ touches the plane

---

1. In what follows we use illumination or illuminance (incoming power density according to visual response) as loosely equivalent to irradiance $E$ (incoming power density). In graphics radiance is $L$ is typically used, related to irradiance for a diffuse emitter and a diffuse reflector with reflectivity $\rho$ as follows: $L = (\rho/\pi)E$.

$z = 0$. We place the origin at $P_{max}$ and the $x$-axis aligned with the projection of the line $L$ (see Figure 2(c)). It now suffices to show the following.

**Lemma 2:** *The restriction of $E(x, y)$ on any line on the plane has a single maximum $t_{max}$ and is decreasing as a function of distance from $t_{max}$.*

In Appendix A we outline the proof of Lemma 2 and show that the cosine of the angle scale factor formed by $OP$ and the normal of the receiver is decreasing as a function of distance from the maximum. We have thus shown that the function of illumination from a disc source on a plane of arbitrary orientation is unimodal. □

In addition, unimodality can be proven for the simple case of a rectangular light source and a receiver plane that is parallel to the plane of the source. This can be shown by taking the antiderivative of the illumination function.

## 3.2 General Characteristics of Illumination Functions

Given the above discussion, there is strong indication that the following conjecture is true.

**Conjecture**: *The illumination functions of planar convex polygonal light sources on arbitrarily oriented receiver planes are unimodal.*

The full proof of the conjecture, as well as finding the largest class of sources and receiver orientations that are unimodal are difficult problems that are subjects of ongoing research. However, the general proof for the disc light source gives us strong evidence that this property holds, since we can see that moderately tight bounds can be constructed from disc light sources for a number of polygonal shapes. Finally, in all the experiments performed by the authors, the illumination functions on surfaces due to convex sources have demonstrated unimodal behaviour. The complicated nature of the functions involved makes establishing these properties quite difficult, as indicated by the above discussion of the simple disc source case. For simplicity, in the following discussion we will consider a one-dimensional cross section of the full illumination function $E(x, y)$, defined as $I(t)$, with the parameter $t$ varying along a line. Such is the graph shown in Figure 1(a). It is important to note that this analysis only gives us a rough idea of how the two-dimensional function behaves.

For polygonal emitters, we have an analytic expression for $E(x, y)$. For a polygonal emitter with $n$ vertices, the illumination at a point $P(x, y)$ is defined analytically by:

$$E(P) = \sum_{i=1}^{n} \gamma_i \cos(\delta_i), \tag{2}$$

where the quantity $\gamma_i$ is the interior angle formed by the point $P$ and the vertices $v_i$ and $v_{i+1}$, and $\delta_i$ is the angle of the normal to the surface defined by $P, v_i, v_{i+1}$ and the normal to the receiver.

What can we conclude about a function's behaviour given unimodality? For a given function $I(t)$, we immediately know that the first derivative has one easily isolated root, at the maximum. The position of this maximum is important, and we call the value of $t$ such that $I(t)$ is maximum, $t_{max}$.

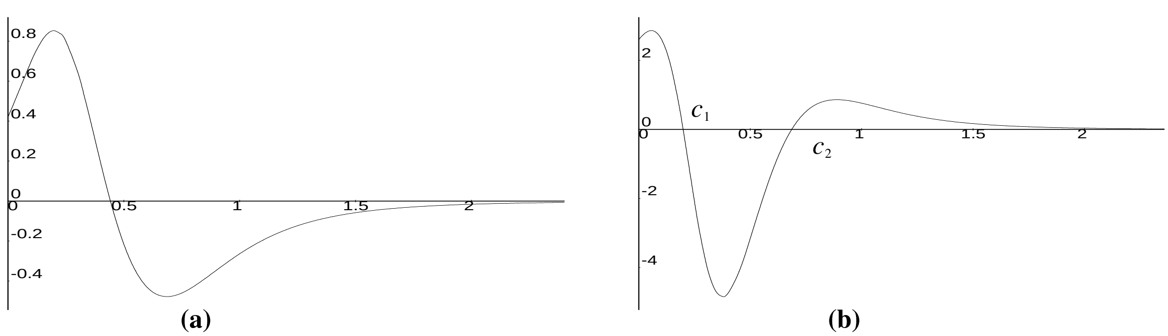



**Figure 3. First and Second Derivatives of a Unimodal Illumination Function on a Linear Subdomain**

Due to the decreasing behaviour of the function with respect to the distance from the maximum, the derivative has one maximum and one minimum (Figure 3(a)). Consequently the second derivative will have exactly two roots. It must be noted that the derivatives for example in Eq. (2) become large very quickly (each differentiation increasing the values of the function by approximately an order of magnitude). In Figure 3(a) the first derivative along the line is shown and in Figure 3(b) the second derivative is shown (both scaled to fit).

One important measure of the smoothness of a function is *curvature*. Curvature is defined as follows:

$$\kappa(t) = \frac{f''}{[1 + (f')^2]} \tag{3}$$

When curvature changes sign, we know that the function changes from convex to concave. To determine these crossings, called inflection points, it is sufficient to find the roots of the second derivative. From now on we call the inflection points, $c_1$ and $c_2$ (see Figure 3(b)). Together with $t_{max}$ they are the *critical points* of the function. For the two dimensional function $E(x, y)$ these points correspond to critical surfaces and are typically of high order in $x$ and $y$.

## 4. Structured Sampling and Polynomial Interpolation

The goal of this research is to develop a sampling strategy and a reconstruction scheme that will allow efficient and accurate representation of the illumination function over a surface. The recognition of the properties of illumination functions discussed previously allows the more effective choice of samples and interpolants.

The choice of sampling strategy is directly linked to the reconstruction scheme used. Piecewise polynomial interpolants have several desirable characteristics: they are computationally efficient, easy to manipulate analytically, and in some cases allow error estimation to be performed. It seems that a sampling scheme that ties in with such interpolants should be non-uniform, so that only as much effort as required is spent in each region of the function. To achieve such non-uniformity, the sample selection must be adaptive, based on an attempt to bound the error.

### 4.1 Representing Illumination with Piecewise Polynomial Interpolants

To maintain the original goal of efficiency, the reconstruction scheme should require minimal computation to determine a value of the desired function. We thus restrict ourselves to linear and quadratic interpolants, since they require a small number of multiplications to evaluate a function at any point.

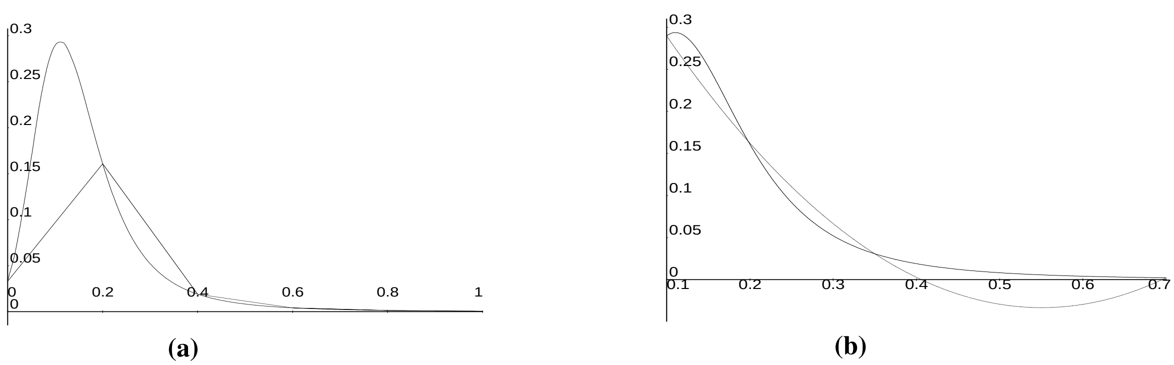



**Figure 4. Undersampling Problems from Uniform Linear and Quadratic Interpolation**

Higher order interpolants can introduce ringing artifacts, and therefore should be used with caution. In previous work, cubic [12] and quadratic [9] interpolants have been used. The following examples indicate some of the possible problems when low sampling densities are used. In Figure 4(a), the function originally depicted in Figure 1(b) is used, over a receiver that is twice as long to the right, to demonstrate the problems more clearly. Using 5 samples, we show the uniform linear interpolation of this function. As can be seen in Figure 4(a), when the function structure is not taken into account, significant errors can result. The maximum value can be missed leading to disastrous results. In the "radiosity gradient" approaches, such initial undersampling will mean that the elements will not be further subdivided, and the maximum illumination will be drastically underestimated. By identifying the maximum $t_{max}$, this artifact can be avoided.

Another interesting artifact is that of over- and under-shooting caused by blind quadratic interpolation (shown in Figure 4(b)). Even though negative lobes can be immediately identified without resorting to the original function structure, non-negative artifacts cannot be consistently identified, unless the regions being examined are guaranteed to be either concave or convex. This can only be accomplished by identifying the critical points.

The use of a polynomial representation is also useful for display. The interpolants can be queried in a ray-casting (as in [9]) or a z-buffer scheme to extract high quality values on a pixel by pixel basis. But we can also resample the interpolants at the vertices of small polygonal elements to be subsequently used in hardware rendering.

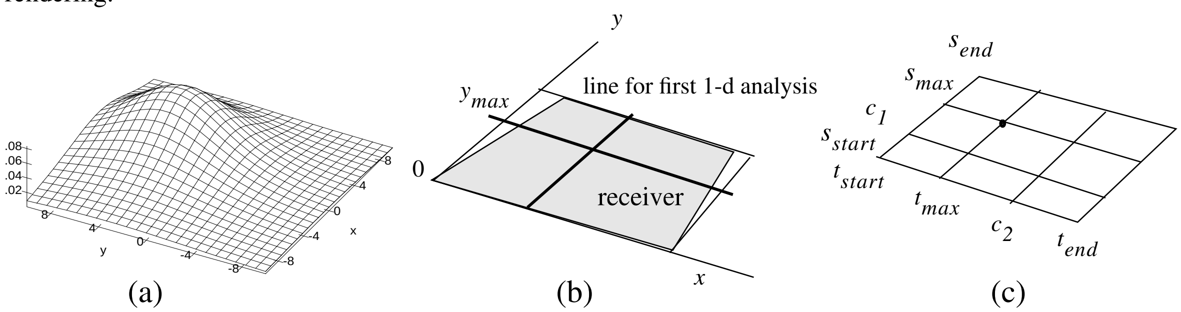



**Figure 5. Finding the Critical Points**

## 4.2 Using One Dimensional Analysis to Generate Tensor Product Interpolants

As noted in Section 3.2, for the full two-dimensional case identifying the critical surfaces and points requires the solution of complicated high order equations. Unimodality however allows us to find the one dimensional maxima in two orthogonal directions, and immediately determine the overall two dimensional maximum. We can thus avoid the computationally expensive solution of the two-dimensional problem. Figure 5(a) shows the illumination function from a rectangular source on a plane that is at an angle to it.

The algorithm initially places an bounding box around the receiver surface, one edge of which is aligned with its longest edge. The two-dimensional local coordinate system is defined as this edge being the *x*-axis, while the first vertex of the bounding box defines the *y*-axis. The midpoints of the lines $y = 0$ and $y = y_{max}$ are found and one-dimensional analysis is performed on the line defined by these two points (see Figure 5(b)). It is first determined whether a maximum exists along this line, and if there is one, it is found numerically. Similar analysis is done in the other dimension (see Figure 5(b)).

We now have two lines, one parallel to the *x*-axis and one parallel to the *y*-axis, on which there either is a maximum, or we have a portion of either the $[t_{start}, t_{max}]$ or $[t_{max}, t_{end}]$ regions (increasing or decreasing respectively). In a similar fashion, the points $c_1, c_2$ are found if they exist. The two lines in the *x* and *y* dimensions, segmented based on the maximum and the curvature, are used to create a set of two-dimensional cells (see Figure 5(c)). The cells are used as a basis for the creation of piecewise bi-linear, bi-quadratic or mixed quadratic/linear tensor product interpolants. In each of the cells, the function is either increasing or decreasing in both *x* and *y*, and either concave or convex. This information is now used to adaptively subdivide the cells. For the same reasons as when finding the critical points, the adaptive subdivision is performed in one dimension, and a tensor product is generated from these final one-dimensional segmentations.

## 5. Algorithms for Adaptive subdivision

Each segmented line can contain the regions $\{ [t_{start}, c_1], [c_1, t_{max}], [t_{max}, c_2], [c_2, t_{end}] \}$. Of these $[c_1, t_{max}]$, $[t_{max}, c_2]$ do not depend on the size of the receiver, while $[t_{start}, c_1]$, $[c_2, t_{end}]$ do. Compare for example Figure 1(b) and Figure 4(a), in which the length of the receiver changed. Due to this distinction in behaviour, different approaches must be taken in dealing with the two kinds of region.

We call the regions $[t_{start}, c_1]$ and $[c_2, t_{end}]$ "tails" of the illumination function. The constant length of the regions $[t_{start}, c_1]$ and $[c_2, t_{end}]$ independent of receiver size allow more assumptions to be made about the function in these sections. Conversely, the tails are typically harder to deal with, since it is difficult to cheaply determine function structure.

To maintain the original goals of efficiency and accuracy, we need a consistent subdivision criterion. Subdivision is thus based on determining an upper bound on the error after a specific subdivision step, and using this bound to determine if subsequent subdivision is necessary. Due to the one-dimensional nature of this analysis, the error bound is actually an upper limit of the error on the one specific line being examined. Due to the unimodal nature of the function, this gives a relatively good approximation to an error bound for the tensor product, but it is still not completely reliable.

Any adaptive subdivision algorithm requires a termination criterion. This is usually a user-supplied parameter. To avoid confusion the tolerance is given in terms of the relative error (e.g. pixels that differ by less than 10%) the user is willing to tolerate, between the approximation and the exact solution.

The adaptive scheme employed is guided by the following general principle. Given a one-dimensional region of the parameter $t$ try to fit a parabola to the function. If an acceptable error bound can be easily determined, stop. If not, attempt to fit a linear approximation within an acceptable error bound, and if that also fails, subdivide. It must be emphasized here that all function values previously computed are reused in the adaptive subdivision steps, and are also used for the final display.

## 5.1 Adaptive Subdivision in the region $[c_1, c_2]$

The algorithm initially is given a region $[a, b]$ which is a subset of either $[c_1, t_{max}]$ or $[t_{max}, c_2]$. Initially a parabolic fit to the function in this region is attempted. To achieve this a third point is required. Define $Q_u(t)$ as the parabola defined by $(a, I(a))$, $(b, I(b))$ and an variable third point $(u, I(u))$. Ideally we should choose $u$ such that $\sqrt{\int_a^b (Q_u(t) - I(t))^2 dt}$ is minimised. Finding this value of $u$ is too expensive, since the integral would have to be numerically evaluated at significant cost.

Fortunately, the illumination function in this region is sufficiently well behaved, so that it is possible to closely approximate the function in one dimension using cubic hermite curves. Using the hermite interpolants allows us to directly solve the following minimisation problem. Define $C(t)$ to be the cubic approximating the function, and $Q_u(t)$ the quadratic defined by $(a, I(a))$, $(b, I(b))$ and $(u, C(u))$. Find $u$ such that

$$F(u) = \sqrt{\int_a^b (Q_u(t) - C(t))^2 dt} \tag{4}$$

is smallest. We call the value of $u$ such that $F(u)$ is minimum $u_m$. If $Q_{u_m}$ is acceptable (see below), then an error bound for this parabola can be determined by finding the maximum value of $B = |Q_{u_m}(t) - C(t)|$. This maximum can be computed directly. It is stressed that the cubics are used only for one dimensional analysis.

The parabolic fit is unacceptable if its maximum is in the interval $[a, b]$, since the interpolant would then not preserve monotonicity in the region. This test can be performed trivially. If this is the case, $u_m$ is set to $(a + b)/2$ and the parabola $Q_{u_m}$ is used.If the parabola is acceptable, the value $B$ is tested, and if it is within tolerance the algorithm stops. If not, a linear fit is attempted, using the previously determined point $u_m$ to generate two linear segments, $L_1$ and $L_2$.

To bound the error incurred by the linear segment $L_i$, the function $|L_i(t) - C(t)|$ is maximised, an operation that again can be performed directly. The value of $t$ for which this is maximum is called $t_m$. In this case the error in the interval due to the linear approximation is bounded by $B = |L(t_m) - C(t_m)|$. For each of the two segments $B$ is compared to the tolerance. If it is less, we are done, and otherwise the segment is subdivided.

## 5.2 A Midpoint Tail Subdivision Scheme

As has been noted, the function behaviour in the tails is less predictable than that in the subsets of $(c_1, c_2)$. As a result the hermite approximation-based techniques cannot be reliably used for the tails. As an alternative, a midpoint subdivision scheme is employed. The algorithm proceeds as follows.

1. Assume that the original interval for subdivision is $[a, b]$. Create the parabola defined by $(a, I(a))$, $(b, I(b))$ and $((a+b)/2, I((a+b)/2))$.
2. Determine if the parabola is an acceptable fit. If so determine the error bounds in both subintervals, due to the parabolic interpolation. If both error bounds are under the acceptable limit, stop. If not, go to Step 3.
3. Determine if the errors from linear interpolants are acceptable in both subintervals. If they are, stop. If one region's bound is above tolerance subdivide and leave the other as is. If not, subdivide and perform Step 1 for both subintervals.

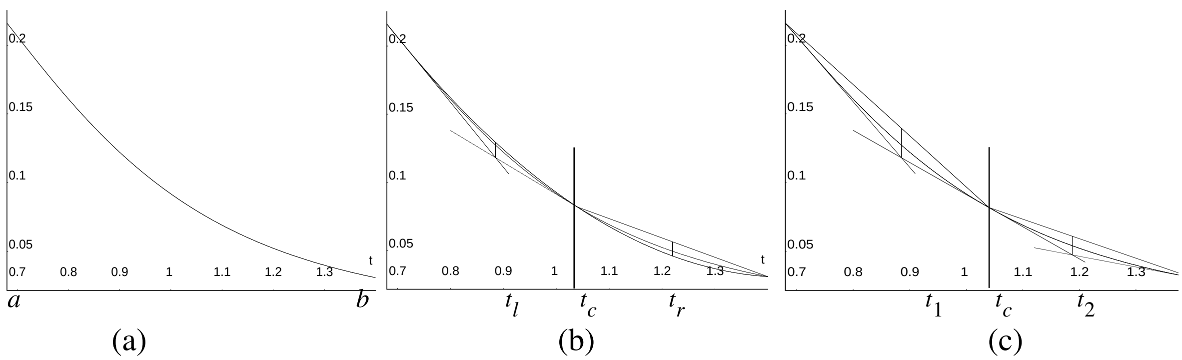



**Figure 6. Bounding the Error for Quadratic and Linear Interpolants**

*Determining Acceptability and Bounding the error for a parabolic interpolant*

Because the regions have been subdivided so that the function in $[a, b]$ is either convex or concave (see for example Figure 6(a)), we can immediately reject any quadratic in which the minimum/maximum lies within the interval $[a, b]$. If this is not the case, the quadratic crosses the function $I(t)$ once in $[a, b]$. Call this crossing point $t_c$ (see Figure 6(b)). The interval $[a, b]$ is now split into two, and we know with certainty that the parabola is completely above or completely below the illumination function in either subregion (see Figure 6(b)). Using this fact, we can immediately determine in which interval the parabola is entirely above the function, and in which it is below, by examining the derivatives of the quadratic and the function at $a$ and $b$.

For the subregion that the parabola is below the illumination function, say $[t_c, b]$ without loss of generality, we can directly compute the point at which the parabola is furthest from the line segment defined by $(t_c, I(t_c))$ and $(b, I(b))$. This situation is shown on the right hand side of Figure 6(b). Call this point $t_r$. An upper bound on the error incurred by the parabola in the subinterval $[t_c, b]$, is $B = |Q(t_r) - L(t_r)|$. The vertical line in the right half of Figure 6(b) shows the magnitude of $B$. For the subregion that is entirely above we employ the bounding method used for linear interpolants, described below. The vertical line in the left half of Figure 6(b) shows the magnitude of the error bound.

*Bounding the error for a linear interpolant*

To bound the error incurred by the line segment $L$ defined by $(a, I(a))$, $(b, I(b))$, we use the derivatives $I'(a)$ and $I'(b)$. Since we know that the function is either convex or concave in this interval, we can create two lines $l_a$ and $l_b$ that have slope determined by $I'(a)$ and $I'(b)$, that will have one intersection in the interval $[a, b]$. These lines, for both left and right hand subintervals, are shown in Figure 6(c). Call this point of intersection $(t_m, I_m)$. The error in $[a, b]$ is bound by the value $B = |I_m - L(t_m)|$. In Figure 6(c), the error bound is shown by a small vertical line for both intervals, where $t_m$ corresponds to $t_1$ and $t_2$ respectively. As we can see this bound is quite tight in most cases.

## 5.3 Computational Expense

The above algorithm is fairly complicated, but an overall effort has been made to avoid much additional computation. Each step of the algorithm described in Section 5.1 only costs as much as the operations required to evaluate cubics and quadratics which is a few multiplies, plus the initial setup cost of the cubic hermite, which is only two extra function evaluations to determine the derivatives at the endpoints. The steps of Section 5.2 however, cost one extra function evaluation for each derivative value computed (maximum of three).

# 6. Implementation and Results

To find the critical points, Brent's minimisation algorithm was used [8]. For a given interval $[a, b]$, it is first determined whether $t_{max}$ is in the interval, by examining the values of the derivative at $a$ and $b$. If it exists, it is found by maximising the illumination function. It must be noted that this expense can be avoided when the maximum can be found geometrically, as is the case for the disc source. In a similar fashion, the points $c_1$ and $c_2$ are found, if necessary, by minimising or maximising the first derivative. The average expense of this iterative procedure over a large number of tests is between 4 and 6 iterations. To avoid numerical instabilities, finite differences are used to find the derivative values.

To compare to the uniform approach, it was necessary to compensate for the expense incurred in finding the critical points and adaptive subdivision. This was achieved by counting the number of function evaluations incurred in the sample placement and the adaptive subdivision. To achieve fair comparison to a uniform solution, a uniform grid is constructed that has the same number of sample points as the structured solution, plus as many additional sample points (within integer limitations) as function evaluations used in the sample placement stage. Both uniform linear and uniform quadratic interpolation were tested. However, the latter resulted in images that had high error. Consequently only uniform linear results are reported.

**Table 1. Test Suites for Rectangle Source and Varying Orientations**

| | **Test 1 Linear** | **Test 1 Struct** | **Test 2 Linear** | **Test 2 Struct** | **Test 3 Linear** | **Test 3 Struct** | **Test 4 Linear** | **Test 4 Struct** | **Test 5 Linear** | **Test 5 Struct** |
|---|---|---|---|---|---|---|---|---|---|---|
| | 9.42 | 6.33 | 19.18 | 5.76 | 4.79 | 5.14 | 4.32 | 0.00 | 2.67 | 2.56 |
| | 12.62 | 5.37 | 19.28 | 3.04 | 4.34 | 2.62 | 4.52 | 0.00 | 2.51 | 0.76 |
| | 13.17 | 3.41 | 18.06 | 2.38 | 4.61 | 2.26 | 4.65 | 0.00 | 1.78 | 0.60 |
| | 13.00 | 3.36 | 15.55 | 2.92 | 6.11 | 3.66 | 4.90 | 0.00 | 0.78 | 0.32 |
| | 6.70 | 0.44 | 13.18 | 5.95 | 5.95 | 2.48 | 5.11 | 0.40 | 0.65 | 0.04 |
| | 8.44 | 0.26 | 6.90 | 2.23 | 7.19 | 0.00 | 3.42 | 0.00 | 0.27 | 0.00 |
| | 7.68 | 0.26 | 8.14 | 2.53 | 7.18 | 0.00 | 4.18 | 1.84 | 9.55 | 3.94 |
| | 6.57 | 0.26 | 2.39 | 3.60 | 5.95 | 2.38 | 5.01 | 2.25 | 6.86 | 2.51 |
| | 7.86 | 0.26 | 0.97 | 4.53 | 4.15 | 3.63 | 7.03 | 0.80 | 3.49 | 2.11 |
| | 8.37 | 0.26 | 0.13 | 5.54 | 4.58 | 2.34 | 7.82 | 2.78 | 2.30 | 2.13 |
| | 6.67 | 0.71 | 0.03 | 6.67 | 4.38 | 2.59 | 7.94 | 4.69 | 2.17 | 2.16 |
| | 13.10 | 2.57 | 0.00 | 7.84 | 4.97 | 5.28 | 8.59 | 2.37 | 2.19 | 2.19 |
| | 13.10 | 3.64 | 0.00 | 8.71 | 2.87 | 1.82 | 4.61 | 4.19 | 2.10 | 2.10 |
| | 9.65 | 4.91 | 0.00 | 2.78 | 0.88 | 0.67 | 1.76 | 3.92 | 2.01 | 2.01 |
| | 9.07 | 6.78 | 0.00 | 2.25 | 0.11 | 0.26 | 0.95 | 4.78 | 1.64 | 1.64 |

**Table 1. Test Suites for Rectangle Source and Varying Orientations**

| | Test 1 Linear | Test 1 Struct | Test 2 Linear | Test 2 Struct | Test 3 Linear | Test 3 Struct | Test 4 Linear | Test 4 Struct | Test 5 Linear | Test 5 Struct |
|---|---|---|---|---|---|---|---|---|---|---|
| % Avg.Err. | 9.69 | 2.59 | 6.92 | 4.45 | 4.54 | 2.34 | 4.99 | 1.87 | 2.73 | 1.67 |
| Std. Dev. | 3.54 | 2.18 | 7.67 | 2.31 | 1.96 | 1.62 | 2.21 | 1.69 | 2.43 | 1.15 |
| Lin/Struct. | | 3.74 | | 1.55 | | 1.94 | | 2.66 | | 1.63 |

## 6.1 Test Results

To evaluate the new algorithm a large number of tests are run of a simple environment, in which only a few parameters vary. We have chosen a rectangular source lighting a simple rectangular receiver. The parameters that can vary are thus distance and orientation of the two polygons. As in previous work by the authors [4], an error metric that measures pixel by pixel image difference from the analytic solution was used. The ratio of the number of pixels that display more than 10% absolute difference from the analytically generated image over the total number of visible pixels gives a percent error metric.

A total of 5 test suites, of 15 images each, were run. Each suite varied one or more of the distance and orientation parameters. In Test 1 the source moves across a horizontally opposing plane in only one direction. In Test 2 the source moves in all three directions. In Test 3 the receiver plane is vertical while a horizontal light source moves in one direction. In Test 4 the receiver plane slowly rotates towards the source and finally in Test 5 a rotated plane moves under the source.

In Table 1, we see the results for a total of 75 tests. The first and last few cases in each suite are typically situations where the tail regions are dominant. Overall we see that the ratio of error from the linear method over the error from the structured method varies between 1.55 to 3.74. This clearly shows that for the same amount of computation, better results are produced. Most of the errors incurred by the structured approach are in the tails and are due to the lack of view dependent considerations in the subdivision approach.

*Frame Consistency in Animation*

Figure 7 shows four frames from an animation sequence with a moving light source. These images are the analytically computed exact solutions. In the colour plates the new method is compared to a uniform solution.

**Figure 7. Analytically Computed Images of Animation**

Colour Plate 1 shows the structured solution results. Notice that in the bright areas the interpolation is of high quality. The Mach bands due to first derivative discontinuities can be seen in the darker areas. Colour Plate 2 shows the grid of samples and the red areas show where the interpolants have overestimated the values of the function. Notice how the grid tracks the movement of the source.

Colour Plate 3 shows the uniform solution. Grid cross-hatching is visible in all images, and the maximum illumination values do not correspond to the real maximum. The cross hatching is inconsistent, since it depends on chance alignment of the grid and the critical points. Colour Plate 4 show the uniform sample grid. Notice how more samples are used to compensate for the expense of determining the critical points and adaptive subdivision. Blue areas are where the interpolant is underestimating the value of the function.

## 7. Conclusions and Future Work

The new method achieves all three goals set forth in the introduction. It is efficient, since the overhead of finding the critical points and adaptively subdividing is definitely worthwhile in terms of the results. It is more accurate than previous methods since adaptive subdivision is based on bounding the error in certain intervals of the illumination function. Finally it results in consistent generation of animated sequences, without the need for user intervention, due to the tracking of important properties of the illumination function (i.e. the position of the maximum and the curvature).

The sampling strategy can be used in any radiosity-based system to improve the quality of sample placement, both for static and moving images. It can be used both in traditional radiosity systems [3], or as a front end for the more sophisticated approaches, such as those described in [9][12]. Use of the interpolants proposed is also a feasible alternative to traditional patch/element representations of radiance over a surface, or the "blind" quadratics used in [9].

In terms of future work, there is much room for improvement and investigation. The use of numerical techniques to identify the critical points is wasteful. Further investigation into geometric techniques will most probably result in faster ways to find $c_1$ and $c_2$. Investigation of the quality/computation trade-off between cubic and quadratic interpolants is an important consideration. Careful analysis of $C^1$ discontinuities should be performed and the results incorporated into the adaptive subdivision algorithm. Triangular interpolants, instead of tensor products may help alleviate this problem. View dependent considerations must also be taken into account as proposed in [4]. The algorithm can be adapted to facilitate this without much modification. Results of the above will significantly improve the quality of the new technique proposed.

In current work the method is being extended to the more general context of a partially occluded environment. The approach is based on computing a discontinuity mesh in the spirit of [11][9] and [6], which results in a face-edge-vertex structure. The algorithm described here is then selectively applied to certain of the faces in light or in penumbra. Multiple sources can also be handled by maintaining a list of polynomial representations of the illumination due to the first few most powerful emitters, and then combining the subsequent less powerful emitters into a simpler interpolant. With these extensions the approach can be used in a general global illumination system.

In conclusion, we have shown that the new algorithm for structured sampling can significantly improve the quality and consistency of images of scenes with simple polygonal light sources. The study of a common class of emitters has allowed the identification of important illumination function properties. These properties are then used to identify critical points, and the use of error bounds for adaptive subdivision results in an efficient, accurate and consistent reconstruction algorithm.

**Acknowledgments**

The authours wish to acknowledge the financial support of the National Science and Engineering Research Council of Canada, the Information Technology Research Centre of Ontario, and the University of Toronto, for this research. Thanks also to Tom Milligan and Moira Minoughan for the photography.

**References**

[1] Baum, Daniel R., Stephen Mann, Kevin P. Smith, and James M. Winget, "Making Radiosity Usable: Automatic Preprocessing and Meshing Techniques for the Generation of Accurate Radiosity Solutions," *ACM Computer Graphics (SIGGRAPH '91 Proceedings)*, vol. 25, no. 4, pp. 51-60, July 1991.

[2] Campbell, A. T. III and Donald S. Fussell, "An Analytic Approach to Illumination with Area Light Sources," *Technical Report TR-91-25, Computer Sci. Dept., University of Texas at Austin*, August 1991.

[3] Cohen, Michael F., Donald P. Greenberg, David S. Immel, and P. J. Brock, "An Efficient Radiosity

Approach for Realistic Image Synthesis," *IEEE Computer Graphics & Applications*, 6:3, March 1986.

[4] Drettakis, George and Eugene L. Fiume, "Concrete Computation of Global Illumination Using Structured Sampling," *3rd Eurographics Workshop on Rendering*, Bristol, UK May 1992.

[5] Hanrahan, Pat, David Salzman, and Larry Aupperle, "A Rapid Hierarchical Radiosity Algorithm," *ACM Computer Graphics (SIGGRAPH '91 Proceedings),* vol. 25, no. 4, pp. 197-206, July 1991.

[6] Heckbert, Paul, "Discontinuity Meshing for Radiosity," *3rd EG Workshop on Rendering*, Bristol, 1992.

[7] Heckbert, Paul, "Radiosity in Flatland," *Proceedings of Eurographics '92*, Vienna, Elsevier Science 1992.

[8] Kahaner David, Cleve Moler and Stephen Nash, Numerical Methods and Software, *Prentice Hall Series in Computational Mathematic*s, New Jersey, 1989.

[9] Lichinski, Dani and Fillipo Tampieri, "Discontinuity Meshing for Accurate Radiosity," *IEEE Computer Graphics & Applications*, vol. 12, no. 6, pp. 25-39, November 1992.

[10] Moon, Parry, The Scientific Basis of Illuminating Engineering, *McGraw-Hill*, 1936.

[11] Nishita, Tomoyuki and Eihachiro Nakamae, "Half-tone Representation of 3-D Objects Illuminated by Area Sources or Polyhedron Sources", IEEE Compsac, 83:237-242 1983.

[12] Salesin, David, Dani Lichinski, and Tony DeRose, "Reconstructing Illumination Functions with Selected Discontinuities," *3rd Eurographics Workshop on Rendering*, Bristol, UK May 1992.

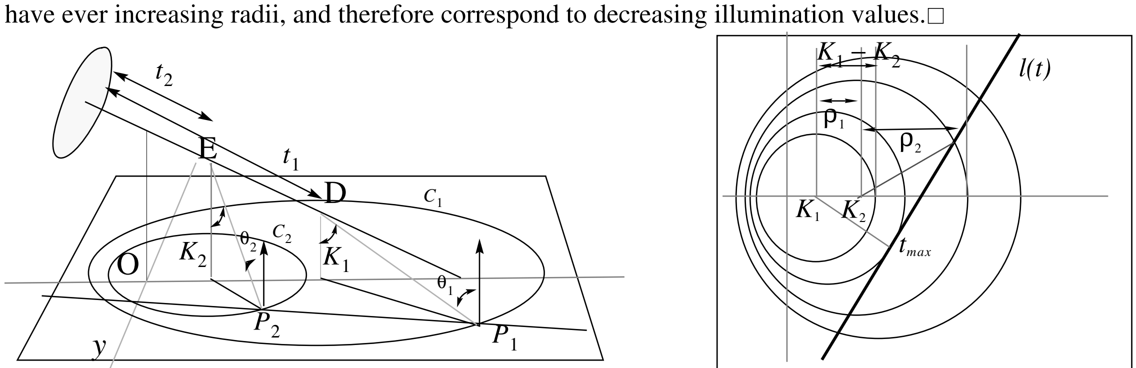
[13] Smits, Brian E., James R. Arvo, and David H. Salesin, "An Importance Driven Radiosity Algorithm," *Computer Graphics (SIGGRAPH '92 Proceedings)*, vol. 26, no. 2, pp. 283-292, July 1992.

## 8. Appendix A

**Lemma 2:** *The restriction of $E(x, y)$ on any line on the plane has a single maximum $t_{max}$ and is decreasing as a function of distance from $t_{max}$.*

Consider the geometry shown in Figure 8(a). Define $K_i$ as the projection of the centre of the sphere $S(t_i)$ onto the *x*-axis. For each sphere there corresponds a circle, say $C_i$, with radius $\rho_i$. It can be shown that for $t_j > t_i$ and consequently $K_j > K_i$, $\rho_j > \rho_i$, and also that $\rho_j > \rho_i + (K_j - K_i)$ (from the equations involved). This can be seen in Figure 8(b) for $K_1$ and $K_2$. Therefore on the plane there is a set of circles that have centres $K_i$ along the *x*-axis, and for $K_j > K_i$, circle $C_j$ entirely encloses $C_i$. This is shown in Figure 8(b).

For any line $l(t)$ crossing this family of circles, there will be only one circle that is tangential to the line, and this point will correspond to the isolux sphere of smallest radius, and therefore the point of maximum illumination $t_{max}$ (see Figure 8(b)). The circles intersected as we move away from the maximum along the line will have ever increasing radii, and therefore correspond to decreasing illumination values.□




(a) (b)

**Figure 8. Proof of Lemma 2**

*The scale factor decreases with distance.* Referring to Figure 8(a), the angle in question is for $P_1$: $\theta_1 = \angle K_1 D P_1$ and for $P_2$: $\theta_2 = \angle K_2 E P_2$, since $D$ and $E$ are the centres of the corresponding isolux spheres. Since we assume that $K_1 > K_2$ and that $\rho_1 = K_1 P_1 > \rho_2 = K_2 P_2$, we can see that:

$$\theta_1 > \theta_2 \Leftrightarrow \cos(\theta_1) < \cos(\theta_2) \quad (5)$$

since $0 < \theta_1, \theta_2 < \pi/2$, and therefore the scale factor preserves the nature of the illumination function that decreases with the distance from the maximum.□