Caustics

Lecture #20: Tuesday, 9 December 1997
Lecturer: Pat Hanrahan
Scribe: Christopher Stolte
Reviewer: Tamara Munzner

In this lecture, we discuss applications of differential geometry within the field of computer graphics. We will see how concepts discussed in earlier lectures can be used to solve problems involving the geometry of optics. Specifically, we will look at Fermat's Principle, rays and wavefronts, and caustics.

1 Fermat's Principle

In geometrical optics, we assume that the wave-like behaviour of light is insignificant and thus model the behaviour of light using rays. Light emitted from a point is assumed to travel along such a ray through space. In an effort to explain the motion through space taken by rays as they pass through various media, Fermat developed his *Principle of Least Action*.

> The path of a light ray connecting two points is the one for which the time of transit, not the length, is a minimum.

At the time that Fermat developed this principle, his justification was more mystical than scientific. His justification can be summarized by the statement that nature is essentially lazy, and these rays are simply doing the least possible work.

We can however develop a more useful formulation of the principle. We know from earlier lectures that the time along a curve through space can be calculated as

$$S(t) = \int dt = \int \frac{ds}{ds/dt}$$

We also know that ds/dt is velocity, which for light is know to be $\frac{ds}{dt} = \frac{c}{\eta}$ where η is the refractive index of the medium. Therefore, we have

$$S(t) = \int \frac{ds}{\frac{c}{\eta(s)}} = \frac{1}{c}\int \eta(s)ds \;\propto \int \eta(s)ds$$

We can thus define the optical path length from one point on a ray to another as the geometric path length weighted by the refractive index of the media. Furthermore, we can now restate Fermat's Principle as

> Light travels along paths of stationary optical path length, where the optical path length is a local maximum or minimum with respect to any small variation in the path.

Determining the path taken by a light ray between two points then becomes a simple matter of optimizing $S(t)$ between the points.

2 Applications of Fermat's Principle

We can make several observations as a result of Fermat's Principle which will prove useful as we explore the realm of geometric optics:

1. In a homogenous medium, light rays are rectilinear. That is, within any medium where the index of refraction is constant, light travels in a straight line.
2. The angle of reflection off of a surface is equal to the angle of incidence. This is the *Law of Reflection*.

We can also make some interesting and useful observations about conic surfaces. Conic surfaces are particulary useful in mirror optics - for example, the design of telescopes. We consider two conjugate points - two points that are perfect images of each other. A salient property of these conjugate points is that the optical path length of all rays connecting them is equal.

Consider a conic surface such as an ellipse. An ellipse is defined as the locus of all points such that the sum of the distances from each point to two fixed points (the foci) is constant, as in Figure 1. The two foci of a mirrored ellipse must then be optically conjugate points. A point source located at one focus must be imaged perfectly at the other focus.

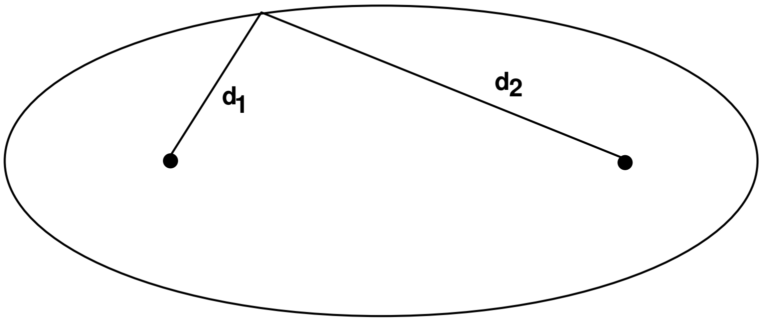

Figure 1: *An ellipse.* The ellipse is defined as the locus of all points such that the sum of the distances from each point to two fixed points is constant: $d_1 + d_2 = c$.

In the case of a parabola, one focus has become infinite. This can be interpreted by saying that an aggregate of rays, parallel to one another and to the axis of a paraboloid after being reflected by the paraboloid, will pass through the focus of the paraboloid. The Newtonian telescope leverages this fact in its design to collect and focus light from distant objects.

In general, a conic surface can be thought of as having two foci and these foci will be optically conjugate points. Figures 2 and 3 illustrate how this property of conic surfaces and geometrical optics has been applied in the design of the Cassegrainian telescope and the Gregorian telescope.

As a side note, a construction called a "Cartesian Oval" is similar to an ellipse, but has weighted distances. That is, rather than being constrained by the equation $d_1 + d_2 = c$, as in Figure 1, the oval is constrained by the equation $n_1 d_1 + n_2 d_2 = c$. The resulting non-elliptical shape will nevertheless have two points of perfect focus.

3 Virtual Light Sources

We now turn our focus to virtual light sources. In traditional ray tracing, a visual ray that encounters a reflective surface is bounced off of that surface and cast in the direction of reflection. Similarly, visual rays are refracted through volumes. Eventually these visual rays reach a non-reflecting surface and the shading calculation is calculated at this point of intersection. This traditional model is depicted in Figure 4.

Although this traditional ray tracing model does allow us to simulate the effect of seeing a scene through a reflective or refractive surface, it does not extend to the simulation of refracted or reflected illumination. In other words, the shading calculation at the point of intersection is limited to the direct components of illumination.

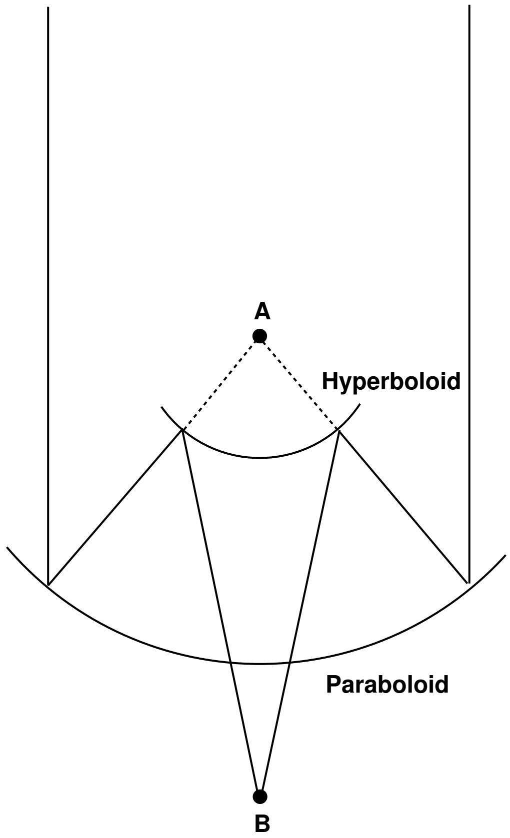

Figure 2: *The Cassegrainian telescope configuration* Point A is where the focus of the paraboloid and the virtual focus of the hyperboloid coincide. Point B is the real focus of the hyperboloid.

We can utilize differential geometry to allow us to solve this more complex problem - determining the reflected illumination at a point on a surface. To do this, we must cast rays from light sources so that they will reflect off of the mirrored surfaces and intersect the point being illuminated. When the light is reflected off the mirrored surfaces it is possible that the light rays may diverge or converge depending on the curvature of the reflective surface at the point of reflection. Thus there are two key problems that must be solved - determining which light rays will intersect the point being illuminated and calculating the proper irradiance at that point. The problem of reflected illumination is depicted in Figure 5.

Finding the paths from the light source to the point $\mathbf{P}$ that reflect off of the mirrored surface is not as complex as might be assumed. A possible path from the light source to the point $\mathbf{P}$ is depicted in Figure 5. The optical path length is

$$d(\mathbf{x}) = \sqrt{(\mathbf{s}-\mathbf{x})^2} + \sqrt{(\mathbf{p}-\mathbf{x})^2}$$

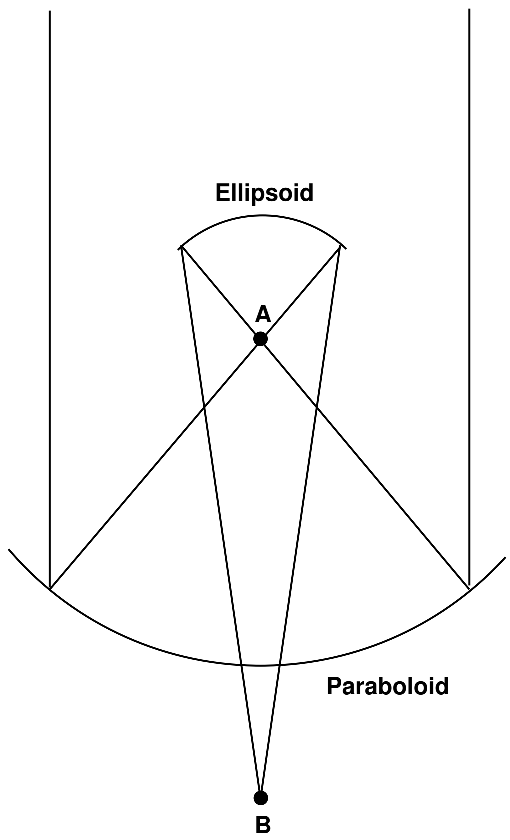

Figure 3: *The Gregorian telescope configuration* Point A is where the focus of the paraboloid and one of the foci of the ellipsoid coincide. Point B is the other focus of the ellipsoid.

According to Fermat's Principle, we want to optimize $d(\mathbf{x})$ in order to determine the path of the light rays. If the mirrored surface is defined by $g(\mathbf{x}) = 0$ then we can optimize $d(\mathbf{x})$ subject to the constraint that $\mathbf{x}$ lie on the surface defined by $g(\mathbf{x})$ using the technique of Lagrange multipliers:

$$\begin{aligned} \nabla d(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) &= 0 \\ g(\mathbf{x}) &= 0 \end{aligned}$$

Solving these equations yields paths of locally extremal length.

Recall from our earlier discussion of conic figures that the two focal points of an ellipse are perfect images of each other - the optical path lengths of all reflected rays connecting

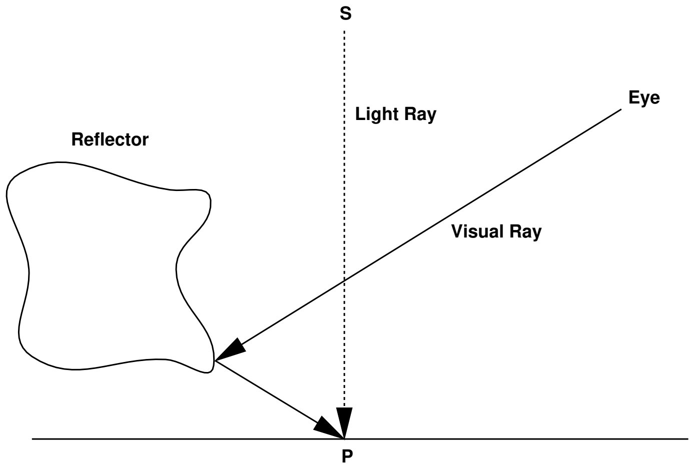

Figure 4: *Reflected Visual Rays*

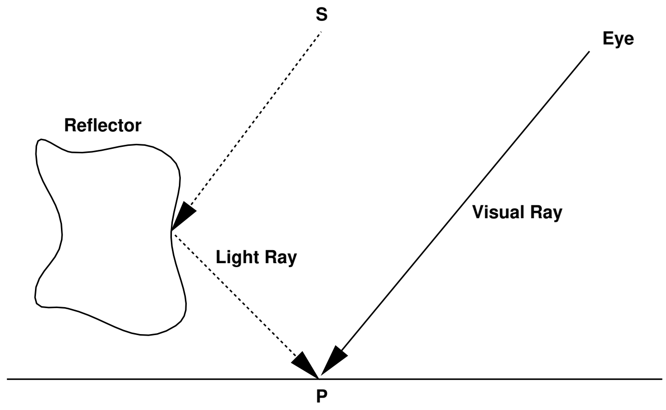

Figure 5: *Reflected Ilumination*

them are equal. If we select $\mathbf{s}$ and $\mathbf{p}$ as our foci of a family of ellipsoids and vary the optical path length, we get a family of confocal ellipsoids.

The system of equations produced by the Lagrange multipliers have a simple geometric interpretation. The extremal points must not only lie on the surface defined by $g(\mathbf{x}) = 0$, but they must also lie on the surface of one of these confocal ellipsoids and the ellipsoid must be tangent to the surface at the point of contact. Figure 6 depicts this geometric

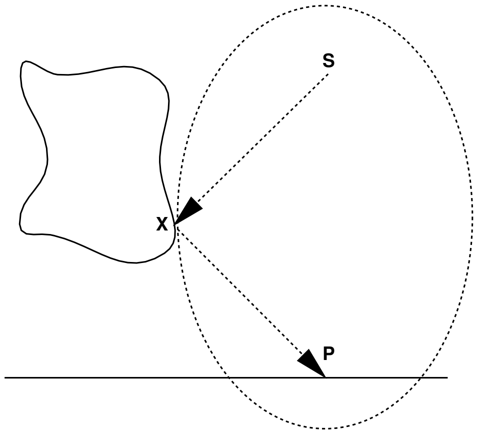

Figure 6: Osculating Ellipsoid

interpretation.

4 Rays and Wavefronts

In order to be able to compute the proper irradiance at a point being illuminated, we will need to determine if the rays of light from the light source are converging or diverging at the point. A given light will be the source of many rays, and the paths of the rays emitted are determined by the following equation:

$$S = \int \eta(s)ds$$

We will define a wavefront W to be the surface defined by the points on each ray at a constant s. Alternatively, the wavefront can be described as the locus of points at a given optical path length. We will not go into the details of wavefront properties, but one important property that should be noted is that the wavefront surfaces are orthogonal to the rays. You can think of wavefronts as isosurfaces in space.

This section is focused on intuitive concepts rather than formal derivations. In this entire discussion, light sources are assumed to be point light sources, although similar concepts and methods can be extended to the area light source case. Figure 7 depicts three simple types of wavefronts: those emitted by a single local point light, those emitted by an infinitely distant point light and a set of converging wavefronts.

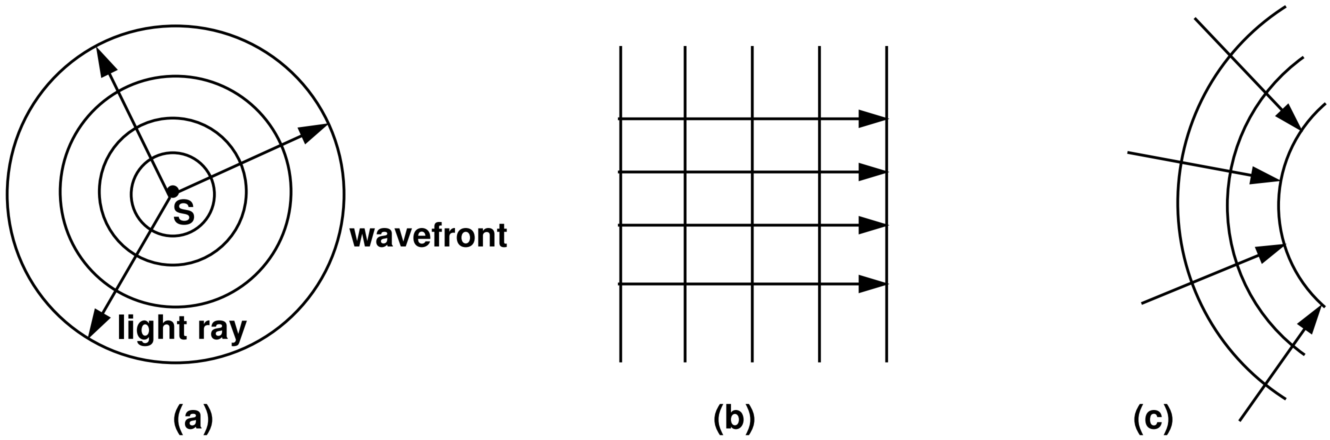

Figure 7: *Three different wavefronts:* (a) those emitted by a single local point light source, (b) those emitted by an infintely distant point light and (c) a set of converging wavefronts.

We are interested in the convergence and divergence of the rays because we need to be able to calculate the intensity of the light at a point being illuminated, and the intensity of the light can be shown to be equal to the radiant power of the light divided by the wavefront area.

Intuitively, this makes sense. Consider a set of rays which represent light moving forward over time, and two wavefronts defined by these rays at different points in time. Furthermore, consider the two patches of area on these wavefronts, shown in Figure 8, which are defined by this set of rays. The situation in the figure is a divergent wavefront, so the area dA' is greater than the area of dA. If the wavefront were converging, the opposite would hold.

We can think of these two patches of area as the ends of a tube containing the set of rays. Although the area of the two patches is different, the total power transmitted through the tube is a constant. Thus the intensity, which can be thought of as the number of rays per unit area, decreases as it passes through the tube. Note again that the intensity would increase in the case of a convergent wavefront.

We can formalize this intuition. We consider the general situation of the neighborhood of a point on a rectilinear ray. There is some orientation of a cutting plane at this point that will yield the maximum radius of curvature r_1, and another orientation of a different cutting plane which will yield the minimum radius of curvature r_2. Furthermore, we know the planes associated with these two radii of curvature are orthogonal from our earlier lectures on differential geometry. These radii of curvature are depicted in Figure 8.

Let dA be this element of area on the wavefront. All rays passing through dA will intersect some subsequent wavefront with area dA'. Let $d\theta_1$, $d\theta_2$ be the elements of angle

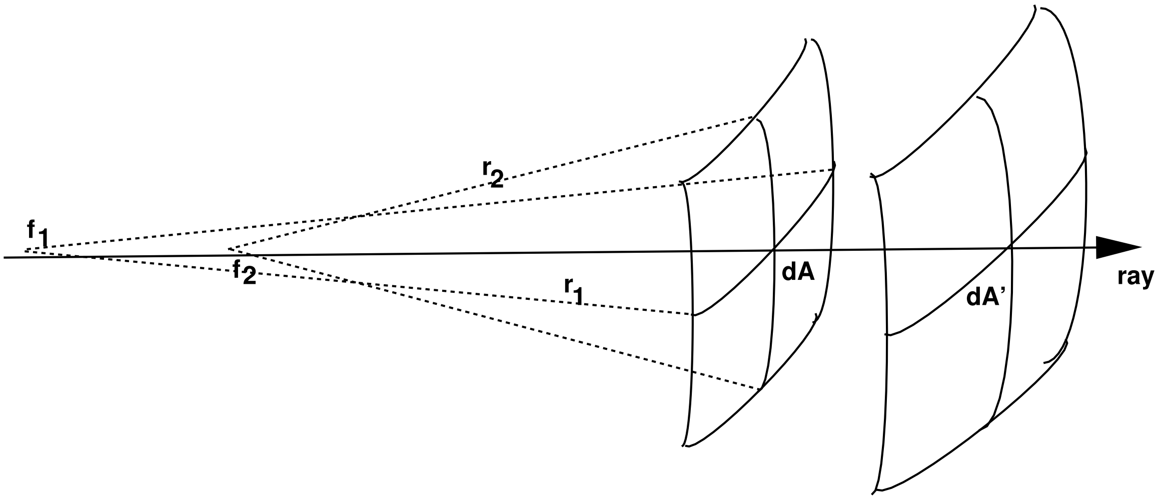

Figure 8: A ray and two small patches of area, dA and dA' on two wavefronts associated with the ray. For small enough patches, the area of the patch is defined by the radi of curvature and the angle subtended at the center of curvatures as $dA = r_1 d\theta_1 r_2 d\theta_2$.

subtended at the centers of curvature by these point areas. Because of the conservation of energy, we must have

$$d\Phi = I' dA' = I dA$$

where $d\Phi$ is power and I is intensity. Therefore

$$\frac{I}{I'} = \frac{dA}{dA'} = \frac{r_1 r_2 d\theta_1 d\theta_2}{r_1' r_2' d\theta_1 d\theta_2} = \frac{r_1 r_2}{r_1' r_2'} = \frac{\kappa'}{\kappa}$$

This illustrates the important point that the intensity is not only related to the area of the wavefront, but also to the inverse of the Gaussian curvature of the wavefront.

We know also that as a wavefront evolves forward according to the principles of optics, that the new wavefront will be an offset surface from the original wavefront. Thus, if the wavefront is diverging we can express the new radii of curvature as

$$\begin{aligned} r_1' &= r_1 + d \\ r_2' &= r_2 + d \end{aligned}$$

Alternatively, if the wavefront is converging, we can express the radii of curvature as

$$\begin{aligned} r_1' &= r_1 - d \\ r_2' &= r_2 - d \end{aligned}$$

Since intensity is inversely proportional to the radi of curvature, this means that at some point there must be infinite brightness. This point of infinite brightness is called a *caustic*, from the dimunitive form of the Greek word for "burning iron".

The caustic is thus the evolute, the locus of the centers of curvature. In the three dimensional case, there will be two caustic surfaces, one for each of the principal directions of curvature. What we colloqially call "caustics" are the curves formed by the intersections of these surfaces with a ground plane or object.

5 Orthotomics

We have seen from above that when a wavefront converges, a caustic is created. We are interested specifically in the case where a point light source shines upon a curved reflector, and the reflected light converges to a caustic. The *orthotomic curve* is an intermediate curve that we will use to find the caustics in this reflected case.

Because of the complexities of the three dimensional case, where the caustic is a curve and there are two caustic surfaces, we will focus our discussion on the orthotomic in the two dimensional case.

We will construct the orthotomic for an arbitrary light source and reflective surface. In doing so, we will show that the orthotomic corresponds to the reflected wavefront.

We construct the orthotomic as follows. Given a source s and a curve c, pick a point x on the curve and find its tangent. Then the locus of reflections of s about such tangents form the orthotomic curve, also known as the secondary caustic. This construction is depicted in Figure 9.

We now explain the construction of the orthotomic more rigorously. In this explanation, let $\mathbf{n}$ be the normal to the curve at point x and $\mathbf{t}$ be the tangent. Let $\mathbf{x}$ be the vector from s to the point x.

We know that if we reflect $\mathbf{x}$ about the tangent at x, this will define the point:

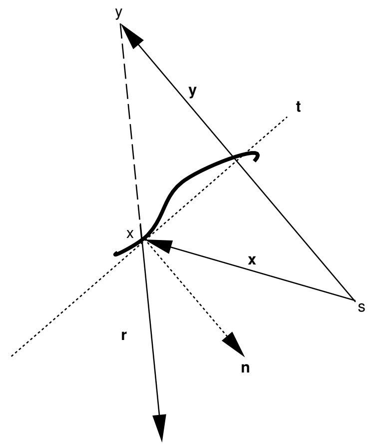

Figure 9: *Construction of the Orthotomic Curve*

$$y = 2[\mathbf{x} \cdot \mathbf{n}]\mathbf{n}$$

Looking at Figure 9 we can understand the formula for the point y. When we reflect the vector $\mathbf{x}$ across the tangent to the curve at x, it defines this new point. We can then note that the projection of $\mathbf{x}$ onto the normal $\mathbf{n}$ multiplied by 2 also gives us this same point. We must multiply by two to account for the fact that we have reflected the vector $\mathbf{x}$ across the tangent. To find $\mathbf{y}'$ we simply differentiate:

$$\begin{aligned}
\mathbf{y}' &= 2[\mathbf{x}' \cdot \mathbf{n}]\mathbf{n} + 2[\mathbf{x} \cdot \mathbf{n}']\mathbf{n} + 2[\mathbf{x} \cdot \mathbf{n}]\mathbf{n}' \\
&= 2[\mathbf{t} \cdot \mathbf{n}]\mathbf{n} - 2\kappa[\mathbf{x} \cdot \mathbf{t}]\mathbf{n} - 2\kappa[\mathbf{x} \cdot \mathbf{n}]\mathbf{t} \\
&= -2\kappa[(\mathbf{x} \cdot \mathbf{t})\mathbf{n} + (\mathbf{x} \cdot \mathbf{n})\mathbf{t}] \\
&\propto (\mathbf{x} \cdot \mathbf{t})\mathbf{n} + (\mathbf{x} \cdot \mathbf{n})\mathbf{t}
\end{aligned}$$

We use identities from the previous lectures: $\mathbf{t} \cdot \mathbf{n} = 0$, $\mathbf{n}' = -\kappa\mathbf{t}$, $\mathbf{x}' = \mathbf{t}$.

Consider the vector $[\mathbf{y} - \mathbf{x}]$. We would like to show that the normal to the orthotomic curve at the point y lies in the same direction as this vector. The normal at the point y must be perpendiculer to the tangent at y (which we will call $\mathbf{t}_y$). If the vector $\mathbf{r}$ (which is the reflection of $\mathbf{x}$ across the tangent to the curve at x) is in the same direction, it too must be perpendiculer to $\mathbf{t}_y$. Since $\mathbf{t}_y = \mathbf{y}'$, it is sufficient to show that $[\mathbf{y} - \mathbf{x}] \cdot \mathbf{y}' = 0$:

$$
\begin{aligned}
[\mathbf{y}-\mathbf{x}]\cdot\mathbf{y}' &\propto [\mathbf{y}-\mathbf{x}]\cdot[(\mathbf{x}\cdot\mathbf{t})\mathbf{n}+(\mathbf{x}\cdot\mathbf{n})\mathbf{t}] \\
&= \mathbf{y}\cdot[(\mathbf{x}\cdot\mathbf{t})\mathbf{n}-(\mathbf{x}\cdot\mathbf{n})\mathbf{t}]-\mathbf{x}\cdot[(\mathbf{x}\cdot\mathbf{t})\mathbf{n}-(\mathbf{x}\cdot\mathbf{n})\mathbf{t}] \\
&= (\mathbf{x}\cdot\mathbf{t})(\mathbf{n}\cdot\mathbf{y})-(\mathbf{x}\cdot\mathbf{n})(\mathbf{t}\cdot\mathbf{y})-(\mathbf{x}\cdot\mathbf{t})(\mathbf{x}\cdot\mathbf{n})+(\mathbf{x}\cdot\mathbf{n})(\mathbf{t}\cdot\mathbf{x}) \\
&= (x\cdot\mathbf{t})(\mathbf{n}\cdot 2(x\cdot\mathbf{n})\mathbf{n})-(x\cdot\mathbf{n})(\mathbf{t}\cdot 2(\mathbf{x}\cdot\mathbf{n})\mathbf{n}) \\
&= 2(\mathbf{x}\cdot\mathbf{t})(\mathbf{n}\cdot\mathbf{n})(\mathbf{x}\cdot\mathbf{n})-2(\mathbf{x}\cdot\mathbf{t})(\mathbf{n}\cdot\mathbf{n})(\mathbf{x}\cdot\mathbf{n}) \\
&= 0
\end{aligned}
$$

We know that the normal to the point y must be perpendicular to the tangent at y, which we calculated above. Since the dot product of the vector $[y-x]$ with this tangent is zero, this vector must lie in the same direction as the normal at y. Furthermore, from the figure and the law of congruent triangles, we can see that the reflected light ray $\mathbf{r}$ from the source must also travel in the direction of $[y-x]$.

Therefore the normal to the orthotomic at y is along the direction of $\mathbf{r}$ and passes through x. In more detail, light from s is reflected by the curve at x, according to the the Law of Reflection. Thus the incident ray and the reflected ray make equal angles on opposite sides of the normal to x. By congruent triangles, the reflected ray is along the line from y to x. From above, this is the normal to y.

It follows then that light rays having the orthotomic as their intial wavefront (i.e light rays starting simultaneously at all points on the orthotomic and then propagating down the normals) are the same as light incident from s and reflected by x. Thus the caustic by reflection of s is the caustic of the orthotomic.

Now, let's sum up intuitively what we have just formally explained. Given a light source and a curved reflector, we want to be able to find the caustics that would be formed. An easy way to compute these caustics is to use the orthotomic curve. The orthotomic curve has the property that its wavefronts will evolve to the same caustics as the wavefronts from the true light source will after being reflected.

6 The Gauss Map

Every point on a surface has some normal $n(u,v)$. The Gauss map is a mapping of every point on a surface to the point on the unit sphere with the same normal. This map is not one-to-one. Figure 10 shows an intuitive sketch of this construction for a small portion of the gauss map. The 3D case is too complicated to draw, so we show the 2D analog.

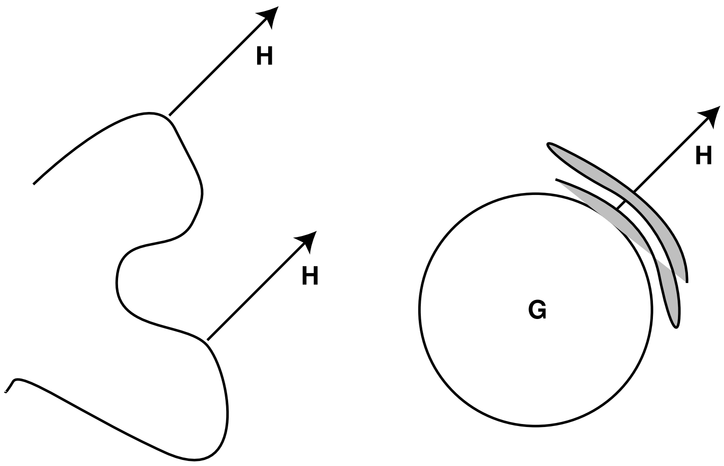

Figure 10: *Construction of the Gauss map:* the multiple normals on the original surface that are in the same direction as the vector N map to the single point on the unit sphere with normal in the direction of N. When the multiplicty of points being mapped to a single point on the unit sphere is greater than one, folds develop in the Gauss map.

The resulting Gauss map may have folds. These folds correspond to inflection points on the original surface, that is, the bottom of a concave valley, the top of a convex hill or a saddle point. At these points, the map which we are tracing out on the unit sphere changes direction because of the change in curvature at the inflection point on the original surface. If the original surface is smooth, the Gauss map will be continuous.

Consider a small patch of area S on the original surface. There will be a corresponding area patch w on the Gauss map. The Gaussian curvature is the differential ratio of the two areas: $\kappa = \lim_{S \to 0} \frac{S}{w}$.

We can formally define the Gauss map:

$$G(x(u, v)) = f(u, v)$$

as a map $G : s = S^2$ from the surface patch S to the unit sphere S^2.

When we are dealing with infinitely distant point light sources, the Gauss map can be used to tell how many virtual lights will be created by a reflective surface. Consider Figure 11. For every position of the viewer and the light source there is a vector H:

$$H = \frac{L + E}{|L + E|}$$

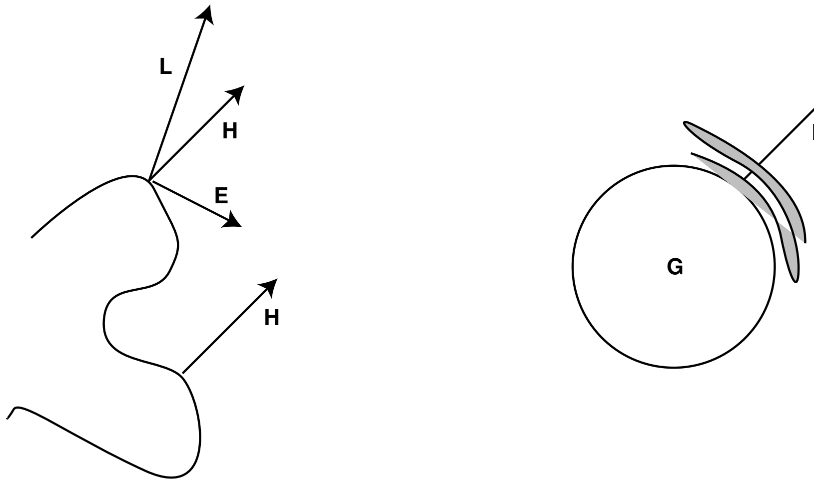

Figure 11: *Specularities* The number of virtual lights is the multiplicity of the points on the Gauss map with normal equal to H.

The number of virtual lights created by a reflective surface is the multiplicity of points on the Gauss map with normal equal to H, that is how many layers exist on the Gauss map.

We are thus interested in the folds created on the Gauss Sphere when the mapping is performed. When the reflective object is deformed or moved, the virtual lights on its surface move, and may be created or destroyed. This creation or destruction of virtual lights occurs at the parabolic points on the Gauss map.

7 Implementation Issues

When trying to implement shading calculations using virtual lights, we can utilize some of the properties we have learned to optimize our calculation. A brief overview of the techniques that can be used when implementing virtual lights is provided here. For a more detailed exposition, see [1].

Most importantly, we can leverage the observation made above that the intensity of light is proportional to the Gaussian curvature of the wavefront associated with that point. Thus, rather than keep track of all of the geometry associated with the wavefronts we can simply track the Gaussian curvature as the wavefronts evolve forward.

Propogating curvature through free space is trivial - we just need to add distance to the radius of curvature. The difficulty lies in efficiently calculating the change in curvature that occurs when the wavefront is reflected off a curved reflector.

We derive the equations necessary to calculate the reflected curvature. Recall the equation giving the directions of the reflected ray:

$$\mathbf{n}^{(r)} = \mathbf{n}^{(i)} + 2\cos i\mathbf{n}^{(s)}$$

In these equations, $\mathbf{n}^{(i)}$ and $\mathbf{n}^{(s)}$ are the normals to the incident wavefront and surface respectively; $\mathbf{n}^{(r)}$ is the normal to the reflected wavefronts. i is the angle of incidence of the rays.

We calculate the vector $\mathbf{u} = \mathbf{n}^{(i)} \times \mathbf{n}^{(s)}$. This vector must be tangent to both the incident wavefront and the surface since it is perpendicular to both normals. We can calculate the curvatures of the incident wavefront in the direction of $\mathbf{u}$ by rotating the principal curvatures using the angle between $\mathbf{u}$ and the line of curvatures and Euler's Formula. The curvatures of the surface in this direction can be computed using the curvature tensor of the surface.

The curvatures of the new wavefronts are computed by taking the directional dervvatives of $\mathbf{n}^{(r)}$ in the direction $\mathbf{u}$. These derivatives can be computed from the formulae for the reflected vectors and the directional derivatives of the normals on the incident wavefront and the surface.

For reflection:

$$\begin{aligned}
\kappa_u^{(r)} &= \kappa_u^{(i)} + 2\cos i\kappa_u^{(s)} \\
\kappa_{uv}^{(r)} &= -\kappa_{uv}^{(i)} - 2\kappa_{uv}^{(s)} \\
\kappa_v^{(r)} &= \kappa_v^{(i)} + (2/\cos i)\kappa_v^{(s)}
\end{aligned}$$

For refraction:

$$\begin{aligned}
\kappa_u^{(t)} &= \eta\kappa_u^{(i)} + \gamma\kappa_u^{(s)} \\
\kappa_{uv}^{(t)} &= \eta\kappa_{uv}^{(i)} + \gamma(\cos i/\cos t)\kappa_{uv}^{(s)} \\
\kappa_v^{(t)} &= \eta\kappa_v^{(i)} + \gamma(\cos i/\cos t)^2\kappa_v^{(s)}
\end{aligned}$$

Remember that the curvature of a plane is 0. Therefore, the curvatures of an outgoing wavefront reflected from a planar surface will be the same as the incoming wavefront (the fact that κ_{uv} switches sign is a result of the change in orientation of the coordinate system due to reflection). This is as expected, since a perfectly reflected wave does not change its shape. Note also that a planar wavefront incident onto a a reflecting surface essentially inherits the curvature of the surface. Thus if the surface is convex, the reflected wavefront will be diverging; whereas if the surface is concave, the wavefront will be converging, eventually forming a caustic.

References

[1] Mitchell,D. and Hanrahan,P., Illumination from Curved Reflectors, *Computer Graphics* 26, 2 (1992), 283-291.

[2] Stavroudis, O. N. *The Optics of Rays, Wavefronts and Caustics*,Academic, 1972.