January 15, 2002

A New Algorithm for Scan Conversion of a General Ellipse

©2002, C. Bond. All rights reserved.

1 Overview

This document contains a derivation of the equations for a general ellipse and implementation details for a new scan conversion algorithm. The algorithm is developed for raster display applications, although its incremental stepping property makes it also suitable for use by plotters.

In the following paragraphs, certain properties of ellipses are developed, analyzed and applied to the drawing problem. Difficult cases are identified and solved by a new algorithm which will be explained in some detail.

The basic incremental algorithms used to plot segments of the ellipse are not particularly original. They are derived from a straightforward analysis of the general ellipse equation using midpoint criteria. However, the published strategies used for moving from one plotted segment to the next will fail in some common circumstances. It is these pathological conditions which are identified and solved in this paper.

Some other important papers on related subjects are: Aken [2], Bresenham [1], and Pitteway [3]. Foley [4] has a detailed discussion and code for the general ellipse, pp.951-961.[1]

2 The General Ellipse

A standard form for general conics which includes ellipses is:

$$Ax^2 + By^2 + 2Cxy + Dx + Ey + F = 0 \tag{1}$$

In this equation, the coefficients of the x and y terms, D and E, represent translation of the ellipse in the x, y plane. This equation can be simplified for

[1]References appear at the end of this paper.

our purposes by translating the center of the ellipse to the origin, eliminating the terms D and E

$$Ax^2 + By^2 + 2Cxy + F = 0 \tag{2}$$

since the process of translation is easily deferred until plot time. This reduces the problem to that of an origin-centered ellipse.

For ellipses with $C = 0$, the axes are aligned with the x, y coordinate axes. Otherwise there is an angle between the axes of the ellipse and the coordinate axes. For any angle, the axes of the ellipse are perpendicular to each other.

Typically, an ellipse would be specified by the length of its major axis, the length of its minor axis, and the angle that the major axis makes with the coordinate system x axis. It is convenient to take the length of the major axis as $2a$ where a is the distance from the center of the ellipse to one end of the major axis. Similarly, the length of the minor axis will be taken as $2b$. The angle from the positive x axis measured counterclockwise to the major axis of the ellipse will be designated θ. This angle will sometimes be referred to as the rotation angle. We will use these conventions in the following treatment.

3 Basic Strategy

For this particular application, we begin by transforming the ellipse equation into a more convenient form. We then calculate values for certain critical locations along the ellipse perimeter, and derive incremental methods for plotting arc segments between these critical points.

Note that when plotting circles an 8-way symmetry reduces the calculation to that of a single octant. For axis aligned ellipses, there is 4-way symmetry, but for ellipses at arbitrary angles we will have to content ourselves with 2-way symmetry.

Figure 1. shows a general ellipse with a, b, and θ identified. Refer to the appendix for the relations between the coefficients A, B, C, F and the given quantities a, b, and θ.

Incremental curve generation is best performed by breaking the curve into segments consisting of arcs whose slopes are confined to a single octant. By this decomposition, it is possible to step along the x or y direction at each iteration, and update the coordinate for the other direction using some control variable.

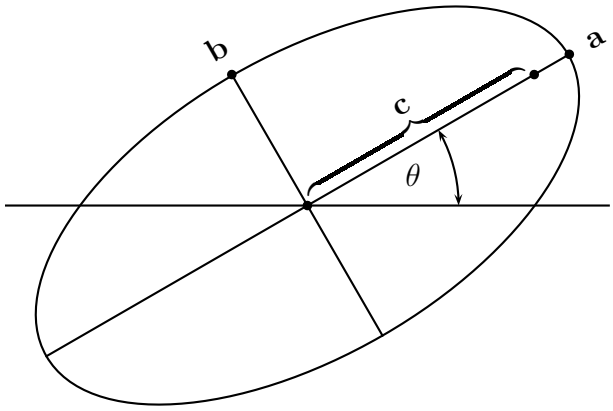


Figure 1. General Ellipse with Labels.

To separate the segments, it is necessary to determine those points on the perimeter of the ellipse which satisfy the following requirements:

- the slope of the tangent is zero,
- the slope of the tangent is infinite,
- the slope of the tangent is ± 1.

By implicit differentiation of Equation (2),

$$\frac{dy}{dx} = -\frac{Ax + Cy}{By + Cx} \tag{3}$$

This equation will be important in determining the specific values of key points on the ellipse. For example, the uppermost and lowermost points on the ellipse can be easily found by observing that the slope of the curve is zero at those points. Hence, from (3) we find

$$\frac{dy}{dx} = 0 = Ax + Cy \tag{4}$$

so that,

$$x = -\frac{C}{A}\, y.$$

Letting $y_{dy/dx=0} = y_t$ and $x_{dy/dx=0} = x_t$ we can substitute into (2) giving,

$$y_t = \pm\sqrt{\frac{AF}{C^2 - AB}} \tag{5}$$

$$x_t = \mp\frac{C}{A}\, y_t \tag{6}$$

for the coordinates of the uppermost and lowermost points of the ellipse. We can also use the expressions developed in the appendix for A, B, C and F to establish that $F = C^2 - AB$ so that we finally have,

$$y_t = \pm\sqrt{A} \tag{7}$$

$$x_t = \mp\frac{C}{\sqrt{A}}. \tag{8}$$

The uppermost point, (x_t, y_t), is associated with the positive square root of the y coordinate. The lowermost point is found by symmetry to be

$$y_b = -y_t \tag{9}$$

$$x_b = -x_t \tag{10}$$

By similar reasoning, for the coordinates of the point where the slope of the tangent is ∞, using Equation (3)

$$By = -Cx$$

so that

$$x = -\frac{B}{C}\, y.$$

Letting $y_{dy/dx=\infty} = y_r$ and $x_{dy/dx=\infty} = x_r$ and substituting in (2),

$$y_r = \pm C\sqrt{\frac{-F}{B(AB - C^2)}} \tag{11}$$

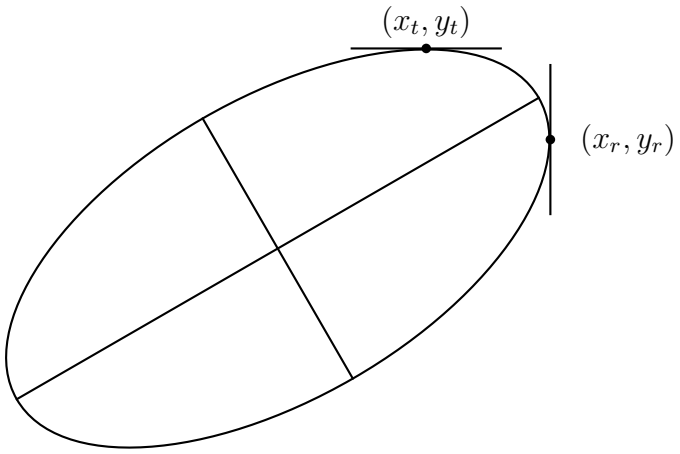
$$x_r = \mp\frac{B}{C}\, y_r \tag{12}$$

for the rightmost point on the ellipse. A final simplification results in

$$y_r = \pm C/\sqrt{B} \tag{13}$$

$$x_r = \mp\sqrt{B} \tag{14}$$

The following figure shows an ellipse with the two critical points x_t, y_t and x_r, y_r identified.

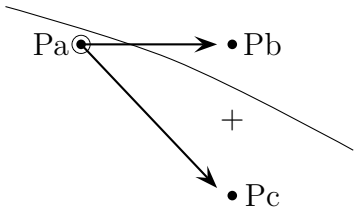


A complete list of the equations for all critical points required (right half of ellipse) is,

$$\begin{aligned}
(x_t, y_t) &= (-\frac{C}{\sqrt{A}}, \sqrt{A}) \\
(x_{tr}, y_{tr}) &= (\frac{B-C}{\sqrt{A+B-2C}}, \frac{A-C}{\sqrt{A+B-2C}}) \\
(x_r, y_r) &= (\sqrt{B}, -\frac{C}{\sqrt{B}}) \\
(x_{br}, y_{br}) &= (\frac{B+C}{\sqrt{A+B+2C}}, -\frac{A+C}{\sqrt{A+B+2C}}) \\
(x_b, y_b) &= (\frac{C}{\sqrt{A}}, -\sqrt{A})
\end{aligned}$$

By restricting the range of each segment so that its slope remains in the same octant, we can reduce the number of choices for the next point to two. This is done by choosing arc segments which are terminated by the critical points where the slope of the tangent is 0, ∞ or ± 1.

Suppose we are plotting clockwise from the topmost point of the ellipse along that portion of the arc where the slope is between 0 and -1 as shown in the following figure.



Here the current point is Pa, and the choice is whether the next point should be Pb or Pc. Clearly, if the curve passes above the midpoint between Pb and Pc, marked with a + sign, point Pb should be plotted. If the curve passes below the midpoint, point Pc should be plotted. In the above case, for example, point Pb is the correct choice.

To guide the plotting process, a control variable is initialized and at each step it is updated. The physical significance of the control variable is that it locates the point midway between the two points which are candidates for the next step. This midpoint value is examined to determine whether it is inside or outside the ellipse. If it is inside, the perimeter is closer to one point. Otherwise it is closer to the other.

4 The Control or Decision Variable

A derivation for the control variable equations will now be given. Only the case for one segment will be considered, because the equations for other segments are derived in a completely analogous manner.

If we have just plotted the ith point along the arc we just described at point Pa, we use the control variable to select the next point from Pb or Pc. The current point is somewhere on the arc from the topmost point of the ellipse to the point where the slope is -1, moving clockwise. The coordinates of the point are x, y. Then the value of the control variable is found by evaluating the ellipse equation at point midway between $x+1, y$ and $x+1, y-1$. Simply substitute $x+1$ and $y-1/2$ into the equation,

$$d = A(x+1)^2 + B(y-1/2)^2 + 2C(x+1)(y-1/2) + F$$

and if d is positive (outside the ellipse) choose Pc, otherwise choose Pb. The rare case in which $d_i = 0$ is a tossup, and is usually decided in favor of the outermost point.

Since stepping along the perimeter is always done incrementally, it is possible to precalculate d_i for the starting point of the arc, and update it with simple quantities at each step. Expressions for the update quantities can be derived by examining the difference between the expression for d_i and d_{i+1}, where d_{i+1} represents one of two possible cases.

At any point, x_i, y_i, in this segment, the value of the expression for the ellipse at the midpoint between Pb and Pc is,

$$\begin{aligned} f(x,y) &= Ax^2 + By^2 + 2Cxy + F \\ d_i &= f(x_i + 1, y_i - 1/2) \\ d_i &= A(x_i+1)^2 + B(y_i - 1/2)^2 + 2C(x_i+1)(y_i - 1/2) + F \\ d_i &= Ax_i^2 + By_i^2 + 2Cx_iy_i + F \\ &\quad +(2A - C)x_i + (2C - B)y_i + A - C + B/4 \end{aligned}$$

If $d_i \leq 0$, the midpoint is on or inside the ellipse and the next point should be Pb. In that case the control variable d_i should be updated to d_{i+1} is follows.

$$\begin{aligned} d_{i+1} &= f(x_{i+1} + 1, y_{i+1} - 1/2) \\ d_{i+1} &= f(x_i + 2, y_i - 1/2) \\ d_{i+1} &= A(x_{i+1}+1)^2 + B(y_{i+1} - 1/2)^2 \\ &\quad +2C(x_{i+1}+1)(y_{i+1} - 1/2) + F \\ d_{i+1} &= d_i + 2Ax_{i+1} + 2Cy_{i+1} + A - C \end{aligned}$$

If $d_i > 0$, the midpoints is on or outside the ellipse and the next point should be Pc.

$$\begin{aligned} d_{i+1} &= f(x_{i+1}+1, y_{i+1}-1/2) \\ d_{i+1} &= f(x_i+2, y_i-3/2) \\ d_{i+1} &= Ax_i^2 + By_i^2 + 2Cx_iy_i + F \\ & \quad (4A-3C)x_i + (4C-3B)y_i + 4A - 6C + 9B/4 \\ d_{i+1} &= d_i + 2(A-C)x_{i+1} + 2(C-B)y_{i+1} + A - C \end{aligned}$$

To summarize the steps required to plot a give segment,

1. initialize the control variables,
2. plot the current best choice for (x_i, y_i),
3. use d_i to determine the next point,
4. update x_i, y_i and d_i appropriately,
5. repeat 2-4 until the chosen segment termination criteria is met.

This process is repeated for each segment. Since an origin centered general ellipse exhibits symmetry across the origin, we may plot two points in step 2 of the above algorithm. The second point is found by reversing the signs of x_i and y_i.

For previously published algorithms, the termination of a segment is signaled by monitoring the slope of the tangent. When that slope passes through ± 1, 0 or ∞ a segment switch takes place.

5 A Plotting Problem and a New Solution

5.1 The Problem

The basic plotting strategy described above is sound — for most ellipses. However, if the aspect ratio (ratio of major to minor axis) of an ellipse is large, sharp corners occur at the extreme points of the ellipse near the ends of the major axis. These corners correspond to regions of rapidly changing slope and it is quite possible for the slope computed at one point to be

totally inappropriate for determining the next step. Here we have a subpixel anomaly which requires special handling.

As stated earlier, previously published plotting strategies plot incrementally from one region to the next, where the boundary between the regions is determined by detecting the point at which the slope equals 0, ∞ or ± 1. At such a decision point the plotting algorithm switches to the next region. Unfortunately, for ellipses with large aspect ratios, it is possible for the curve to completely turn through one or more regions within a single pixel distance. When this happens the updated control variable is incorrect and the plot may wander far from its intended course. This renders the plotting strategy useless.

5.2 The Solution

Our solution is to precompute the coordinates of the critical points and store them for reference. These points locate the boundaries of the arc segments in a definitive and unambiguous manner. Note that for narrow ellipses several critical points may cluster in a region consisting of a few adjacent pixels, or even a single pixel. This is the condition that defeats slope controlled algorithms.

A stable plotting algorithm can be devised, however, by simply plotting each arc from its start to the next critical point, rather than plotting until some slope criteria is met. When the next critical point is within a single pixel distance, we simply plot it and advance to the next region. If the critical point following this is identical to or adjacent to the current critical point, we skip the next segment entirely and just plot the critical point. Similarly, if any successive critical point is within a single pixel distance, we plot it and skip that segment. No 'breakaway' conditions occur with this strategy and all points are plotted correctly. In short, instead of using the slope to control the plotting process, we use the proximity of the current location to a well-defined critical point.

Ellipses with large aspect ratio will have two clusters of critical points, one at each end of the major axis. Broad regions of relative flatness join these clusters. It is important to ensure that when plotting along one flat side we never step across the major axis and wander into the wrong region. This condition can be prevented by forming the equation for the line representing the major axis and testing pixels to verify that they stay on the correct side.

Appendix

The purpose of this appendix is to determine the coefficients, A, B, C, and F of Equation (2) in terms of the major axis, $2a$, the minor axis, $2b$, and the rotation angle, θ.

An ellipse can be defined as the locus of points the sum of whose distances from two fixed points (called foci) is constant. Other definitions are possible, but this is a fundamental property which will be used to derive our equations.

Let the major axis length be $2a$ and the minor axis length $2b$. The distance from the to center to either focus is c, with a subscript sometimes used when it is necessary to distinguish one focus from the other. The total, constant distance from a focus to any point on the ellipse and on to the other focus is s. We will place the center at the origin. Note that the foci lie on the major axis.

With these conventions, the distance from the center of the ellipse to the end of the major axis is a and the distance to the end of the minor axis is b. The distance from the center to either focus $c_{1,2}$ is

$$c = \sqrt{a^2 - b^2}. \tag{15}$$

This follows from the property that the distance from c_1 to b to c_2 must be the same as the distance from c_1 to the point at the opposite extreme a to c_2. That is,

$$2\sqrt{b^2 + c^2} = c + a + (a - c) = 2a. \tag{16}$$

The tilt angle, θ is the angle between the major axis a and the x axis of the coordinate system. The location of the rightmost focus, for example, is then

$$x_c = c\cos(\theta), \tag{17}$$
$$y_c = c\sin(\theta). \tag{18}$$

We can now form an equation involving these relations and satisfying the requirements for an ellipse,

$$\sqrt{(x - x_c)^2 + (y - y_c)^2} + \sqrt{(x + x_c)^2 + (y + y_c)^2} = 2a \tag{19}$$

which expresses the property that the sum of the distances from any point on the perimeter, x, y, to the foci is a constant s.

To determine the values of A, B, C, and F in Equation (2), we must expand (19) in powers of x and y.

The urge to clear out radicals, nurtured by past experience, is irresistible. Begin by rearranging Equation (19), squaring and simplifying.

$$\sqrt{(x+x_c)^2+(y+y_c)^2} = 2a - \sqrt{(x-x_c)^2+(y-y_c)^2} \quad (20)$$

$$xx_c + yy_c = a^2 - a\sqrt{(x-x_c)^2+(y-y_c)^2} \quad (21)$$

Again rearranging, squaring and simplifying,

$$a\sqrt{(x-x_c)^2+(y-y_c)^2} = a^2 - (xx_c + yy_c) \quad (22)$$

$$a^2x^2 + a^2(x_c^2+y_c^2) + a^2y^2 = a^4 + x^2x_c^2 + 2x_cy_cxy + y^2y_c^2 \quad (23)$$

Collecting terms in powers of x and y yields,

$$(a^2-x_c^2)x^2 + (a^2-y_c^2)y^2 - 2x_cy_cxy + a^2(x_c^2+y_c^2-a^2) = 0.$$

Substituting c^2 for $x_c^2+y_c^2$,

$$(a^2-x_c^2)x^2 + (a^2-y_c^2)y^2 - 2x_cy_cxy + a^2(c^2-a^2) = 0,$$

and, finally, replacing c^2-a^2 with $-b^2$, we have

$$(a^2-x_c^2)x^2 + (a^2-y_c^2)y^2 - 2x_cy_cxy - a^2b^2 = 0. \quad (24)$$

From Equations (2) and (24),

$$A = a^2 - x_c^2 \quad (25)$$

$$B = a^2 - y_c^2 \quad (26)$$

$$C = -x_cy_c \quad (27)$$

$$F = -a^2b^2 \quad (28)$$

References

[1] J. E. Bresenham, "A Linear Algorithm for Incremental Digital Display of Digital Arcs", *Comm. ACM,* Vol 20, No. 2, Feb. 1977, pp. 100-106.

[2] Jerry R. Van Aken, "Efficient Ellipse-Drawing Algorithm", *Texas Instruments Incorporated.*

[3] M. Pitteway, "Algorithm for Drawing Ellipses or Hyperbolae with a Digital Plotter", *C*omputer J., Vol 10, No. 3, Nov. 1967, pp.282-289.

[4] James D. Foley, *et al.*,"Computer Graphics, Principles and Practice," *Addison-Wesley Publishing Company*, 1997.