Draft, May, 1997

MATHEMATICS AND ART

Marc Frantz
Department of Mathematical Sciences
IUPUI
402 North Blackford Street
Indianapolis, Indiana 46202-3216
mfrantz@math.iupui.edu

This project was supported, in part,
by the
National Science Foundation
Opinions expressed are those of the authors
and not necessarily those of the Foundation

Linear Perspective

1. Basic Results

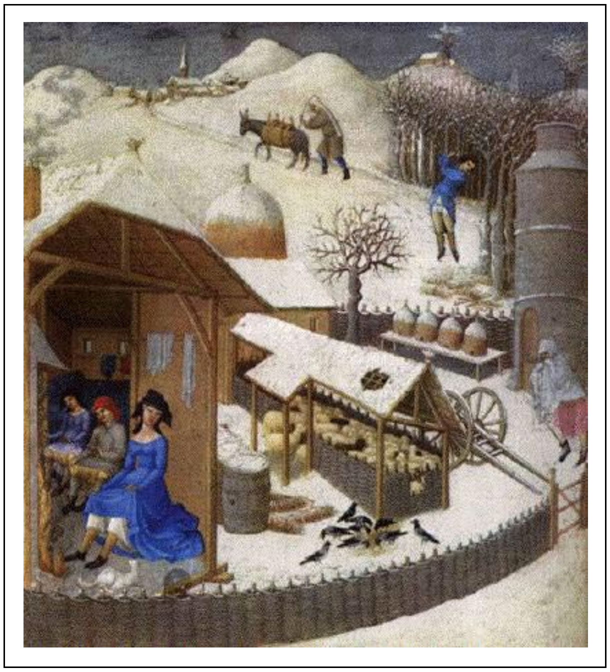


Figure 1(a).

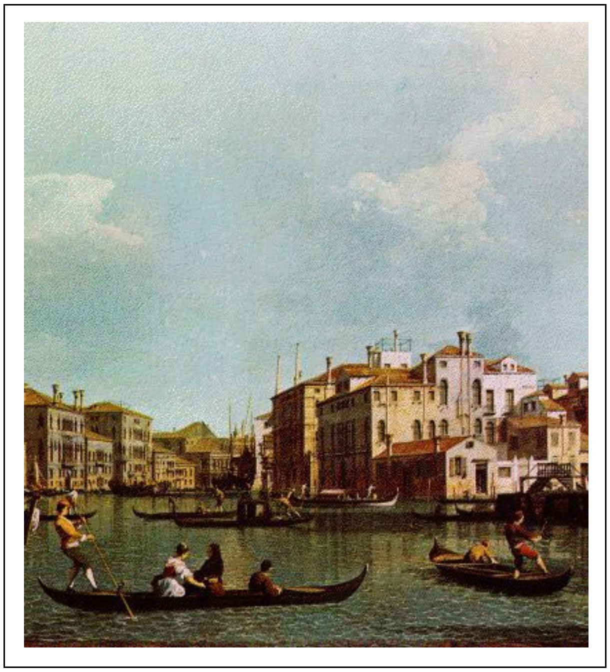


Figure 1(b).

Take a good look at the images in Figure 1. Figure 1(a) is a detail from an illuminated manuscript called the *Très Riches Heures,* painted by the Limbourg brothers, Paul, Hermann and Jean, in the fifteenth century. Figure 1(b) is a detail from an eighteenth century oil painting by the Italian artist Canaletto (Giovanni Antonio Canal), entitled *View of the Grand Canal towards the Pallazzo Contarini dagli Scrigni.* Of all the things one might notice about these works, we would like to focus on one specific aspect, namely the illusion of depth in three-dimensional space, which is far more effective and believable in the painting on the right. Given that the artists in either case were considered very competent in their time, it would appear that something happened in Europe between the fifteenth century and the eighteenth century that added a powerful new tool to the artist's toolbox. The "something" was the Renaissance, a period of vigorous and creative activity in the arts and sciences. The powerful new tool was a mathematical technique called "linear perspective," or just "perspective."

An illusion of depth is not necessarily the most important aspect—or even a necessary aspect—of a successful painting. But whenever the effect has been needed, it has been found that a knowledge of mathematics is an indispensable tool in achieving it. In fact, this is why we have chosen to compare two paintings from Europe, for nowhere was the effect of perspective upon art more dramatic. Moreover, the use of mathematical perspective is perhaps more important today than ever before, if one considers modern creative media such as filmmaking, computer graphics, and virtual reality that have been made possible by advances in technology.

Figure 2. Computer-aided perspective helps bring to life a landscape with dinosaurs in *Jurassic Park*.

The idea behind the concept of "linear perspective" is illustrated in Figure 3. We imagine a viewer using only one eye, with the eye located at $E(0, 0, -d)$ in a three dimensional coordinate system, looking in the direction of the positive z-axis. The point $P(x, y, z)$ on the vase in the figure is visible to the viewer, and light reflected from that point travels in a straight line to the viewer's eye, piercing the "picture plane" $z = 0$ at the point $P'(x', y', 0)$. We think of the picture plane as the surface of a canvas on which we wish to paint a realistic picture of the vase; thus the point P' on the canvas should be painted as a small dot of the same color as the light reflected from the point P on the vase, so that the viewer will see the same thing from that direction that he or she would see if the vase were actually there.

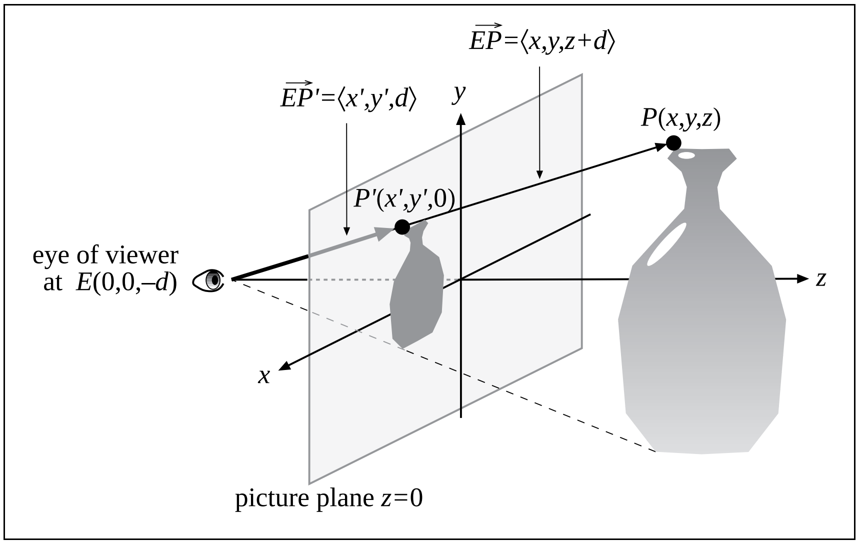

$\overrightarrow{EP}=\langle x,y,z+d\rangle$
$\overrightarrow{EP'}=\langle x',y',d\rangle$
y
$P(x,y,z)$
$P'(x',y',0)$
eye of viewer
at $E(0,0,-d)$
z
x
picture plane $z=0$

Figure 3. P' is the perspective image of P.

The point P' is called the *perspective image* (or sometimes the *central projection*) of P. If we identify the vase with the set of all visible points on its surface, then the perspective image of the vase (represented by the small vase image in the figure) is the set of perspective images of those visible points.

We only need the two coordinates (x', y') to locate a point in the picture plane. From now on a pair (x', y') of primed coordinates will indicate a point in the picture plane (the coordinates of this point in space would be $(x', y', 0)$). To determine x' and y' in Figure 3, notice that the vectors $\overrightarrow{EP'} = \langle x', y', d\rangle$ and $\overrightarrow{EP} = \langle x, y, z + d\rangle$ are parallel, and hence there is a positive number t such that

$$t\langle x', y', d\rangle = \langle x, y, z + d\rangle.$$

Equating components and solving for u, v, and t, we get

$$x' = x/t, \qquad y' = y/t, \qquad t = (z + d)/d.$$

Substituting for t in the expressions for x' and y', we get

$$x' = \frac{dx}{z + d} \qquad \text{and} \qquad y' = \frac{dy}{z + d}. \tag{1}$$

Thus, if we know the coordinates (x, y, z) of a point on the vase—or any other object—and if we know the distance d from the viewer's eye to the picture plane, we can determine the coordinates (x', y') of the corresponding perspective image in the picture plane using the formulas in (1).

Example 1. Assume we have the setup in Figure 3, with the viewer 3 units from the picture plane. If $P(2, 4, 5)$ is a point on an object we wish to paint, find the picture plane coordinates coordinates (x', y') of the perspective image of P.

Solution. We have $d = 3$, $x = 2$, $y = 4$, and $z = 5$. Thus, by (1),

$$x' = \frac{(3)(2)}{(5) + (3)} = \frac{6}{8} = \frac{3}{4} \qquad \text{and} \qquad y' = \frac{(3)(4)}{(5) + (3)} = \frac{12}{8} = \frac{3}{2}.$$

Notice in Figure 3 that we have put the vase on the opposite side of the picture plane from the viewer; that is, the z-coordinates of all the points on the vase are positive. Although it is possible to do the same kind of perspective treatment when the object is between the viewer and the picture plane (in which case the image gets larger than the object), we shall assume from now on that any object we wish to make a picture of consists of points whose z-coordinates are positive. Thus, in a sense, the image will always be smaller than the object. (We say "in a sense" because we have not been clear about the meaning of "smaller." To be more precise, if P_1 and P_2 are two points on the object, then the distance between their corresponding perspective images P'_1 and P'_2 will be smaller than the distance between P_1 and P_2; Exercise # gives you some hints and asks you to prove this.)

Of course, we would like to do problems that are more interesting than Example 1, but on the other hand, we don't want to knock ourselves out doing lots of tedious computations. To obtain some nice shortcuts that will help us do interesting drawings, we will derive important drawing rules from (1). To start with, we would like to say that the perspective image of a straight line is also a straight line. Unfortunately, that's not quite true. For example, a line that goes through the viewer's eye, such as the line through E and P in Figure 3, would be represented by a single point in the picture plane (which makes sense, because a thin pencil lead, seen end-on, looks like a dot, not a line segment). However, the following theorem gives us some help in this regard. For convenience, we stick to our rule that any object we might want to depict consists of points with positive z-coordinates.

Theorem 1. *If S is a subset of a straight line L in space (S might be a line segment, a point, a ray, etc.), and if all the points of S have positive z-coordinates, then the perspective image S' of S is a subset of a straight line $Ax' + By' = C$ in the picture plane.*

Proof. Let us assume that L is given parametrically by

$$x = x_0 + at, \qquad y = y_0 + bt, \qquad z = z_0 + ct,$$

where $-\infty < t < \infty$ and x_0, a, y_0, b, z_0, c are constants. Now let P be any point of S. Then $P = (x_0 + at, y_0 + bt, z_0 + ct)$ for some value of t, and by (1), the perspective image $P' = (x', y', 0)$ of P satisfies

$$x' = \frac{d(x_0 + at)}{(z_0 + ct) + d} \qquad \text{and} \qquad y' = \frac{d(y_0 + bt)}{(z_0 + ct) + d}.$$

(Since $z = z_0 + ct$ is positive, we are not dividing by zero.) If we define the constants A, B, and C by

$$A = bz_0 + bd - cy_0, \qquad B = cx_0 - ad - az_0, \qquad C = d(bx_0 - ay_0),$$

it is straightforward (but tedious) to check that $Ax' + By' = C$. Since the point P of S was arbitrary, the theorem is proved. ■

Although it's unfortunate that we had to say something technical instead of "The perspective image of a straight line is also a straight line," there are some useful drawing tricks related to Theorem 1. One useful related fact is that the perspective image of a line segment $\overline{P_1P_2}$ is not only a subset of a straight line, it is (not surprisingly) either a point or a line segment $\overline{P'_1P'_2}$, where P'_1 and P'_2 are the corresponding perspective images of P_1 and P_2 (Figure 4).

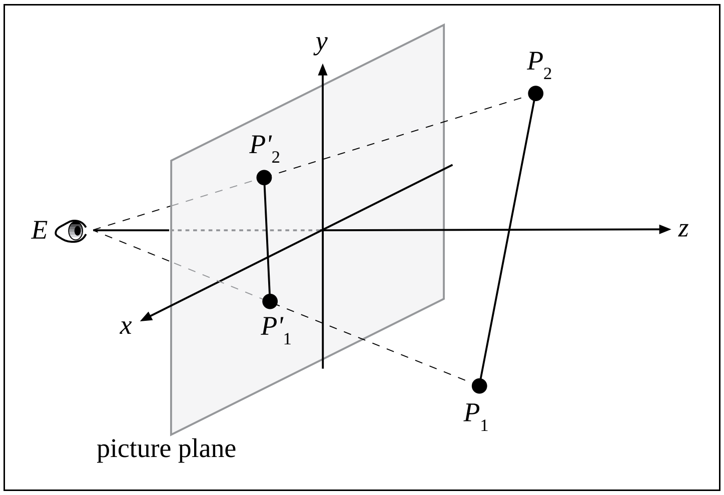


Figure 4. The perspective image of $\overline{P_1P_2}$ is $\overline{P_1'P_2'}$.

Instead of proving this fact, let's use it to do

Example 2. The visible corners of a cube in space are

$M(4,-2,3)$, $N(6,-2,3)$, $P(4,-4,3)$, $Q(6,-4,3)$, $R(4,-2,5)$, $S(6,-2,5)$, $T(4,-4,5)$.

If the viewer is located at $E(0,0,-3)$, make a line drawing to show the cube the way the viewer sees it.

Solution. Our first step is to use (1) to determine the $x'y'$-coordinates of the corresponding perspective images M', N', P', Q', R', S', T'. Since the z-coordinate of E is -3, we have $d = 3$, and we can, for example, compute the $x'y'$-coordinates of T' as follows:

$$x' = \frac{(3)(4)}{(5)+(3)} = \frac{12}{8} = \frac{3}{2} \qquad \text{and} \qquad y' = \frac{(3)(-4)}{(5)+(3)} = \frac{-12}{8} = -\frac{3}{2}.$$

Proceeding in this fashion, we can specify each perspective image with an ordered pair (x', y') and get the points

$$M'(2,-1),\ N'(3,-1),\ P'(2,-2),\ Q'(3,-2),\ R'\left(\frac{3}{2},-\frac{3}{4}\right),\ S'\left(\frac{9}{4},-\frac{3}{4}\right),\ T'\left(\frac{3}{2},-\frac{3}{2}\right).$$

Now the edges of the cube are line segments in space, and by our comment above, their perspective images are line segments in the picture plane. Thus, to make a line drawing of the cube, we plot the primed points that represent the corners, and connect appropriate corners with line segments (Figure 5).

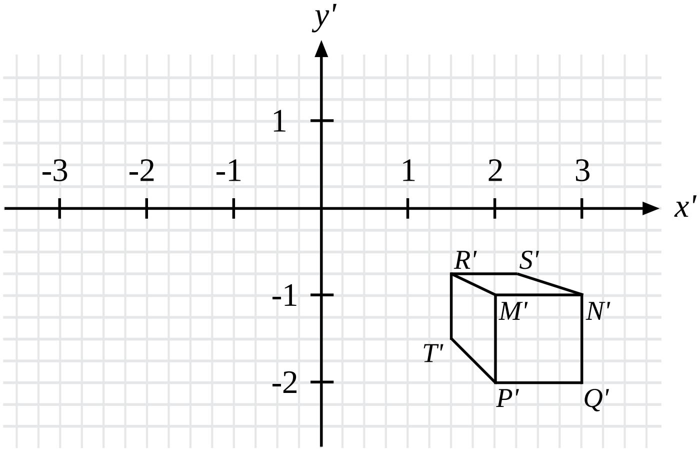


Figure 5. A cube in perspective.

Although Figure 5 is not exactly a masterpiece, we can make some interesting observations about it. First however, test your grasp of the geometry involved by answering these questions:

(i) There is one corner point $U(x, y, z)$ of the cube which is not visible. What are its coordinates? Deduce them by thinking about the coordinates of M, N, P, Q, R, S, and T.

(ii) In Figure 5, use a pencil to indicate where you think the perspective image U' of U would be if the cube were transparent. What are its picture plane coordinates (x', y')?

(iii) Check your answers to questions (i) and (ii) for consistency by applying (1) to your coordinates for U to determine x' and y'.

Now let's make some observations about Figure 5. We know that the object depicted there is a cube, and hence angles such as $\angle QPM$ and $\angle TPM$ are in reality right angles. However, if we look at the corresponding angles $\angle Q'P'M'$ and $\angle T'P'M'$ in Figure 5, we see that $\angle Q'P'M'$ is a right angle, while $\angle T'P'M'$ is not. Similarly, each face of the cube is in reality a 2×2 square, but the only visible face of the cube whose image is a square is the face $MNPQ$, whose image is the 1×1 square $M'N'P'Q'$ in Figure 5. If you answered Questions (i) and (ii) correctly, you found that the hidden corner U had coordinates $(6, -4, 5)$ and its image U' had picture plane coordinates $(9/4, -3/2)$. Thus the image $R'S'T'U'$ of the hidden "back face" $RSTU$ is also a square, and its dimensions are $3/4 \times 3/4$. In other words, the two faces $RSTU$ and $MNPQ$ of the cube have perspective images $R'S'T'U'$ and $M'N'P'Q'$ that are miniature, but undistorted, versions of the original square faces. It also happens that these two faces of the cube are parallel to the picture plane $z = 0$. This can be checked by noticing that the points R, S, T, U all have a z-coordinate of 5, and the points M, N, P, Q all have a z-coordinate of 3. Thus R, S, T, U belong to the plane $z = 5$, while M, N, P, Q belong to the plane $z = 3$.

This state of affairs is not an accident; in fact, it reflects a more general principle which artists have taken advantage of for hundreds of years, namely, that any shape which lies in a plane parallel to the picture plane will have a perspective image that is an exact, undistorted miniature of the original (Figure 6). Such shapes are relatively easy to draw, because they don't have to be "distorted" mathematically to give an illusion of depth.

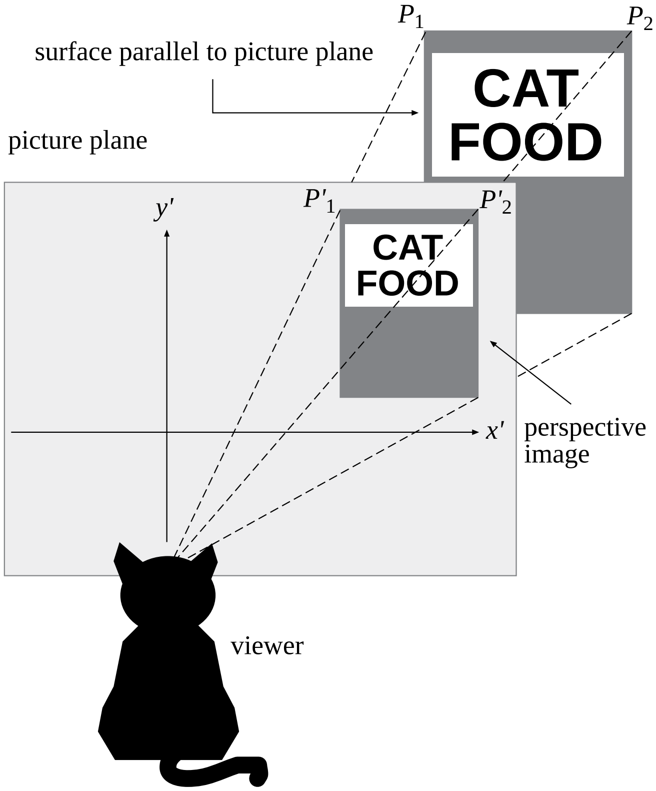


Figure 6. Since the cat food label is parallel to the picture plane, its perspective image is undistorted.

You probably have a pretty good idea of what we mean by an "undistorted miniature" by looking at Figure 6, but we should be precise about the idea, since it turns out to be very convenient for artists. As a first step in making things more precise, we present Theorem 2: while reading it, think of S as the set of points on the cat food label in Figure 6, and think of P_1 and P_2 as any two points on the label.

Theorem 2. *Let S be a set of points in a plane $z = z_0$ $(z_0 > 0)$ parallel to the picture plane, and let the viewer be located at $E(0,0,-d)$ with $d > 0$. Then there exists a constant λ, with $0 < \lambda < 1$, such that for any two points $P_1(x_1, y_1, z_0)$ and $P_2(x_2, y_2, z_0)$ in S, their corresponding perspective images $P_1'(x_1', y_1', 0)$ and $P_2'(x_2', y_2', 0)$ are exactly λ times as far apart as P_1 and P_2, and the vector $\overrightarrow{P_1'P_2'}$ is parallel to the vector $\overrightarrow{P_1P_2}$. In other words, $\overrightarrow{P_1'P_2'} = \lambda\overrightarrow{P_1P_2}$.*

Proof. From (1) we can compute

$$x_1' = \frac{dx_1}{z_0 + d}, \qquad y_1' = \frac{dy_1}{z_0 + d}, \qquad x_2' = \frac{dx_2}{z_0 + d}, \qquad y_2' = \frac{dy_2}{z_0 + d}.$$

If we define the constant λ by $\lambda = d/(z_0 + d)$, then $0 < \lambda < 1$ since d and z_0 are positive constants, and we can rewrite the above equations as

$$x_1' = \lambda x_1, \qquad y_1' = \lambda y_1, \qquad x_2' = \lambda x_2, \qquad y_2' = \lambda y_2.$$

Thus

$$\overrightarrow{P_1'P_2'} = \langle x_2' - x_1', y_2' - y_1', 0\rangle = \langle \lambda x_2 - \lambda x_1, \lambda y_2 - \lambda y_1, 0\rangle = \lambda\langle x_2 - x_1, y_2 - y_1, z_0 - z_0\rangle = \lambda\overrightarrow{P_1P_2}.$$

■

To grasp the meaning of Theorem 2, suppose that the viewer is 3 units from the picture plane (that is, let $d = 3$), and consider an object, say, an equilateral triangle with sides of length 2, that lies in the plane $z = 4$. Then the number λ in Theorem 2 is given by $\lambda = d/(z_0 + d) = 3/(4 + 3) = 3/7$. Theorem 2 says that the images of any two points in the plane $z = 4$ will be 3/7 as far apart as the actual points; thus the image of each side of the equilateral triangle will be a line segment of length $(3/7) \cdot 2 = 6/7$. It follows that the image of the equilateral triangle is also an equilateral triangle, and the angles are preserved (their images are still 60° angles). It's not hard to prove in general that any angle in a plane $z = z_0$ will have an image with the same degree measure (Exercise #). What we have said so far means that "shapes are preserved." But what about orientation? That is, is it possible that the image of an equilateral triangle is an upside down version of the original triangle, or tilted at some strange angle? The answer is no, because the vector $\overrightarrow{P_1P_2}$ connecting any two points P_1, P_2 (such as two vertices of the triangle) in the plane $z = 4$ will have a perspective image $\overrightarrow{P_1'P_2'} = (3/7)\overrightarrow{P_1P_2}$ that is parallel to it, so the image is not rotated in any way.

This non-distortion effect can be seen in Figure 7. In this picture, we are looking at a picture plane that is vertical with respect to the (locally flat) earth. Several buildings on the right side of the painting have walls that are parallel to the picture plane, and hence their images, as well as the images of their doors and windows, are undistorted. As we mentioned earlier, this makes it easier for the artist to render these parts of the painting, and indeed many paintings and drawings of buildings feature walls that are parallel to the picture plane. However, convenience is not the only reason for drawing buildings in this way. Remember that most paintings are done on a rectangular canvas, and hence objects in the painting that have a rectangular shape will help organize the composition and keep it in harmony with its border. (Several lines that parallel the edges of the canvas have been indicated in the figure.) Even if the face of a building is not parallel to the picture plane, the edges of the buildings—which are vertical lines in space—will each lie in some plane that is parallel to the picture plane, and hence by Theorem 2 the images of the lines will also be vertical; this adds still more lines that are parallel to edges of the canvas, such as the two vertical lines indicated at the left side of the picture. One compositional use of such lines is illustrated by the vertical edge of the tall building that can just barely be seen at the far right edge of the canvas. As the viewer's eye scans the rooftops from left to right, the eye tends to be led off the edge of the painting. This vertical line serves to "stop the eye" and bring the viewer's attention back to the center of the painting. This little trick is a direct application of the rule $\overrightarrow{P_1'P_2'} = \lambda\overrightarrow{P_1P_2}$ derived in Theorem 2: The vector $\overrightarrow{P_1P_2}$ is the actual vertical edge of the building, and the line segment in the painting is the vector $\overrightarrow{P_1'P_2'}$, which is also vertical, and λ times smaller.

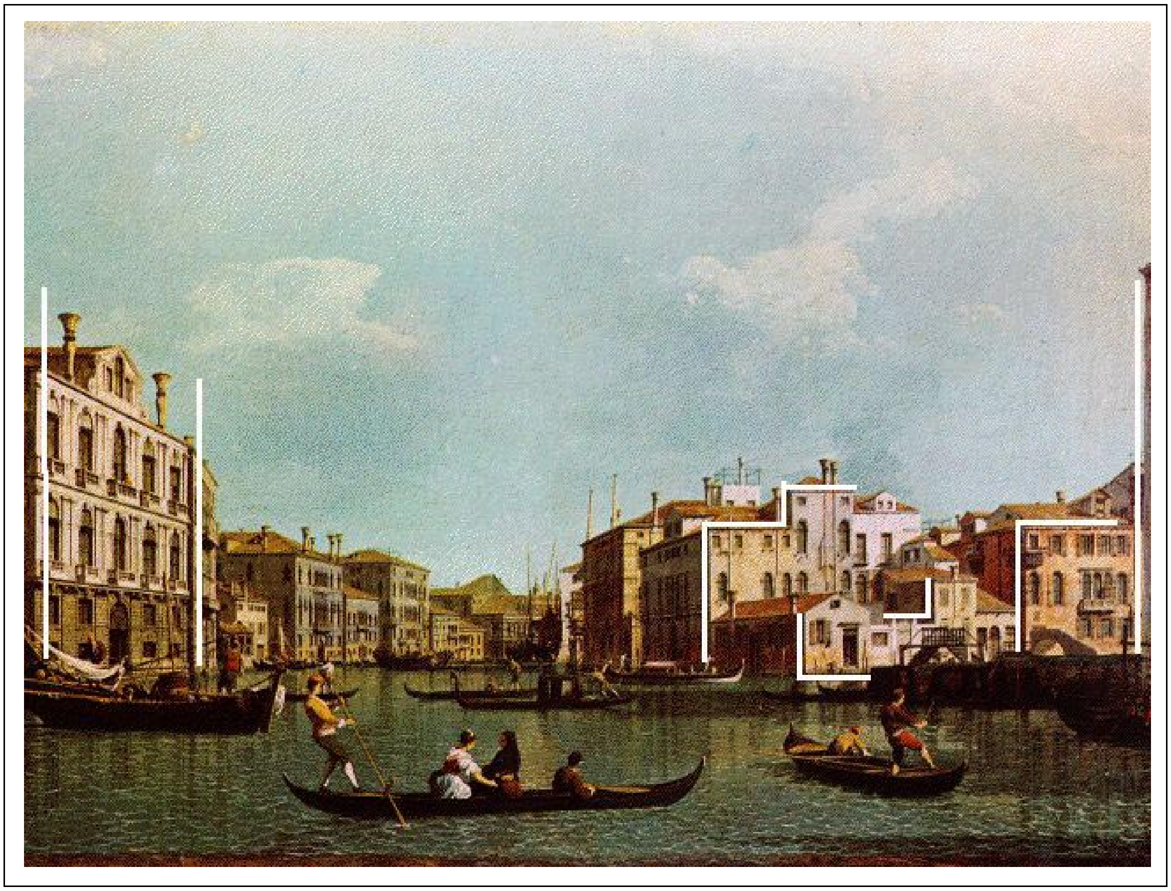


Figure 7. A larger view of Canaletto's painting in Figure 1a. Horizontal and vertical lines help organize the composition and relate the objects in the painting to the border of the canvas.

Besides the non-distortion phenomenon, there is another important idea we can grasp by looking at the cube from Figure 5, which has been reproduced in Figure 8 without all the labels. Take a look at the cube we drew. Does it really look like a cube? If you're looking at it from a typical reading distance, it probably seems too elongated in the direction of your line of sight—more like the shape of dumpster or a cedar chest, than a die or a sugar cube.

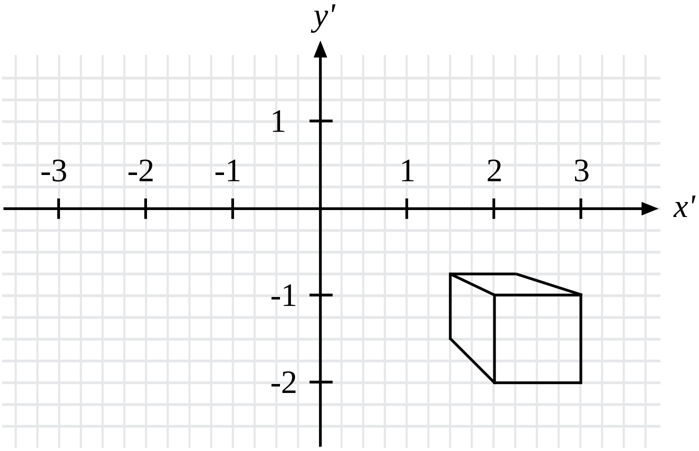


Figure 8. The drawing from Figure 5. At a natural viewing distance (as opposed to the correct 3 units) the cube looks too long in the direction of the line of sight.

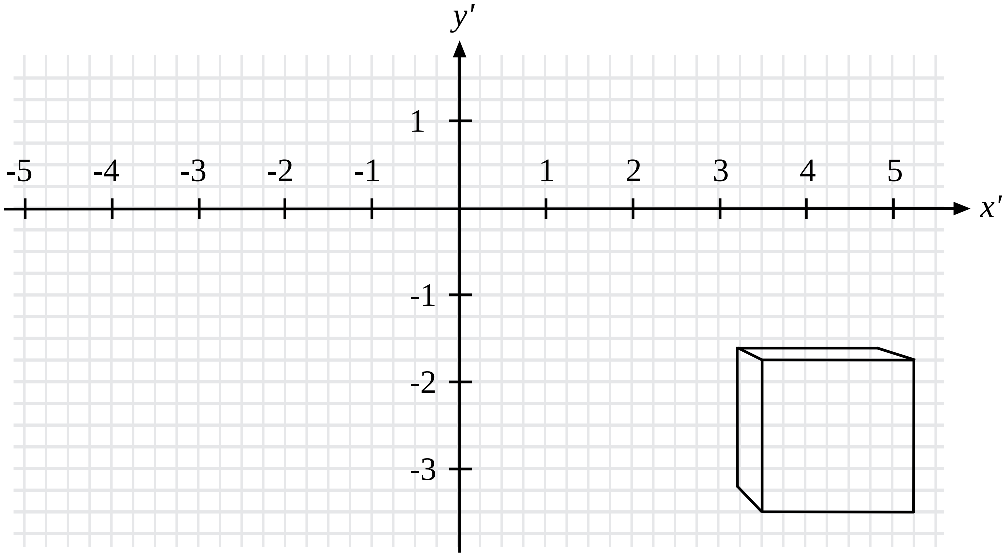


Figure 9. The same cube and picture plane, with the correct viewing distance increased to a more natural 21 units.

There's a good reason–or perhaps we should say a bad reason—for this. Remember that the correct viewing distance from the picture plane is $d = 3$ units. If you hold a scrap of paper up to the x'-axis, mark off 3 units, and use the paper to place your eye 3 units from the origin, your nose will almost touch the page! At this distance you won't be able to focus properly, but if you glance off to the side towards the image, you should be able to notice that it now looks like a (blurry) cube. Thus, even though our drawing is *mathematically* correct, it's not a good drawing, because during the planning stage we failed to take into account the actual finished size of the drawing and the natural distance at which a viewer would want to look at the picture. If we use the same cube and picture plane, but change the viewing distance to a more reasonable value of 21 units (that is, leave everything the same as in Example 2, except move the viewer back to $E(0, 0, -21)$) we get the drawing in Figure 9, which looks much more cube-like from a normal viewing distance.

Now you may argue that the only thing wrong with Figure 8 is that the drawing is too small. After all, if it were so big that the units were, say, in feet, then the viewing distance would be 3 feet, which is far enough away that one could at least focus on the image. To see if this approach works, we have made an enlarged detail of Figure 8 in Figure 10.

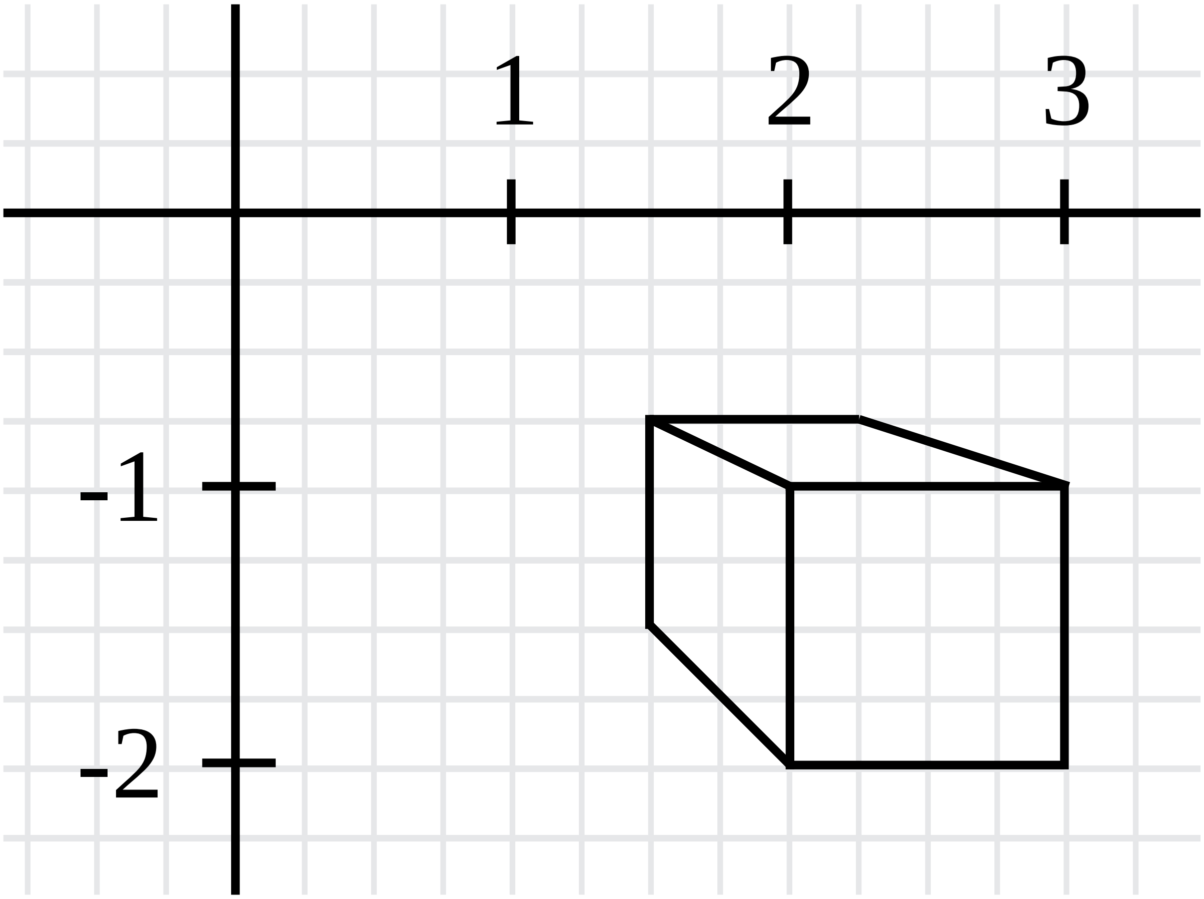


Figure 10. Enlarged detail of Figure 8.

Of course, we couldn't make it big enough so that the viewing distance is 3 feet, but you should be able to focus on the image with your eye 3 units from the origin. At this point, however, a new problem becomes evident: the image of the box is so far from the center of view (the origin) that it is awkward to look that far off to the side. In fact, no matter how much we enlarge the picture, anyone viewing from 3 units away from the origin (measured perpendicularly from the page) would have to look off to the side at more than a 45° angle to see the lower right corner of the box. (Prove it!) It is generally accepted that a drawing or painting done in perspective should not require the viewer to have a "cone of view" of more than about 60°; that is, the viewer should not have to look away from the center of view more than about 30° in any direction to see an object in the painting, so our drawing is not that easily fixed.

Before going further, let's highlight some important points.

- A drawing done in perspective should be planned so that the correct viewing distance is practical and comfortable for the viewer (unless, of course, unusual effects are deliberately intended).
- With the same qualification, the viewer should not be required to look away from the center of view more than about 30°.
- The enlightened viewer (You!) should be aware that there is exactly one correct point in space from which to view a work done in perspective.

Is it possible for the "enlightened viewer" to determine the correct viewing location? In many cases, the answer is yes, and we'll be doing that later. Right now, have some fun with Figure 10.

While looking at Figure 10, close one eye and look at the picture from arm's length. Now move the page closer until your eye is about 3 units directly in front of the origin. Move the page back and forth like this and notice that, as we said earlier, the box seems elongated in the direction of your line of sight when viewed from far away, and more cube-like when viewed closer in. From *any* viewing distance, the image is a perfectly good drawing of a rectangular box with its sides parallel to the coordinate planes (we'll show why this is true later). But, due to the elongation and shortening effects we have just described, it is a correct representation of a *cube* only when viewed from the correct viewing distance. Hence what we perceive from too far away (or too close) is not just a reduction (or an enlargement) of the image the artist wants us to see, it is a *distortion.* This applies to any work done in perspective. We'll prove it for images of rectangular boxes with their faces parallel to the coordinate planes, and then easily generalize it to any kind of image.

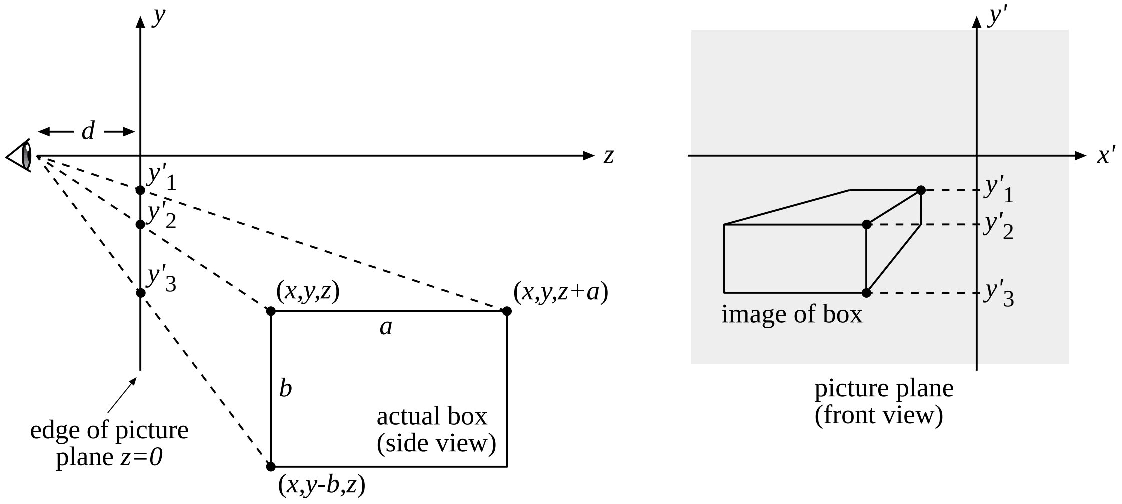


Figure 11.

Theorem 3. *Let B be a rectangular box in space, with positive z-coordinates and sides parallel to the coordinate planes. Suppose also that at least one face of B not parallel to the picture plane is visible to a viewer located at $(0, 0, -d)$. If this face has dimensions $a \times b$, where a is the length in the z-direction, then the viewing distance d is directly proportional to the "shape ratio" a/b .*

Proof. For definiteness, let us suppose that a face of the box parallel to the yz-plane is visible to the viewer (Figure 11). Choose corner points $(x, y, z + a)$, (x, y, z),$(x, y - b, z)$ of this face as in Figure 11, and let y'_1, y'_2, y'_3 be the y'-coordinates of their respective images in the picture plane. From (1) we have

$$y'_1 = \frac{dy}{(z+a)+d}, \qquad y'_2 = \frac{dy}{z+d}, \qquad y'_3 = \frac{d(y-b)}{z+d}.$$

Solving these equations for a, $(z + d)$, and b, respectively, we get

$$a = \frac{dy}{y'_1} - (z+d), \qquad (z+d) = \frac{dy}{y'_2}, \qquad b = y - \frac{y'_3}{d}(z+d).$$

Substituting dy/y'_2 for $(z + d)$ in the solutions for a and b and simplifying, we get

$$a = dy\frac{y'_2 - y'_1}{y'_1 y'_2} \qquad \text{and} \qquad b = y\frac{y'_2 - y'_3}{y'_2},$$

and hence

$$\frac{a}{b} = d\frac{y'_2 - y'_1}{y'_1(y'_2 - y'_3)},$$

or equivalently,

$$d = \frac{y'_1(y'_2 - y'_3)}{y'_2 - y'_1}\left(\frac{a}{b}\right).$$

Thus d is directly proportional to a/b. The case in which a face of the box is parallel to the xz-plane can be handled similarly. ■

Here is what Theorem 3 tells us. Suppose we are given a drawing of a rectangular box like that described in Figure 11. The drawing has already been done, so all the $x'y'$-coordinates are fixed, and hence the quantity $\frac{y'_1(y'_2 - y'_3)}{y'_2 - y'_1}$ is a constant, which we will call K. We can then write

$$d = K\left(\frac{a}{b}\right). \tag{2}$$

(In the case in where we consider a visible face of the box parallel to the xz-plane, we still get an equation of this form, but the constant K depends on x'-coordinates instead of y'-coordinates.) As we mentioned earlier, from *any* viewing distance, the drawing can be considered an accurate image of *some* box with sides parallel to the coordinate planes. But Equation (2) says that if d is doubled, then the shape ratio a/b of a particular side of the box must be doubled also. In other words, if you view the picture at twice the correct distance, you'll perceive a box which is twice as long in the z-direction as the box the artist intended you to see. Similarly, if you change from d to $d/3$ (i.e., if you view too closely), then the box you think you see will have the shape ratio of that same side changed to $(a/b)/3$; that is, the box will appear too short in the z-direction by a factor of $1/3$. Thus Theorem 3 is just a precise quantitative description of what you experience when you view Figure 10 from different distances. You've already noticed that the box looks too elongated when you view its image from a distance of 6 units instead of the correct 3 units; now you know exactly *how much* it's elongated. The mathematics has quantified your experience!

It is easy to generalize the implications of Theorem 3 to arbitrary shapes. Suppose you have, say, a sculpture of a human head you want to draw in perspective. Suppose further that the sculpture is made from tiny cubes—such as salt crystals, for instance—with their faces parallel to the coordinate planes you are using for your perspective setup. Theorem 3 says that if your drawing is viewed from too far away (or too close), then each cube will appear too elongated (respectively, too compressed) in the direction of the z-axis. But if this applies to each cube, it also applies to the entire sculpture, so the entire image will be appear distorted if viewed from the wrong distance. The funny thing is, we see this kind of distortion so often that we don't even notice it. For example, hold your fist out at arm's length with the thumb side towards you and look at it. If you were to take a photograph of what you're seeing and then view the photo from very close up, you might say to yourself, "Well, I guess this is what my fist looks like from close up." But it isn't! To verify this, move your fist so close to your eye that the bottom knuckle of your thumb almost touches your eye. (Remove your glasses if necessary.) Notice that your fist almost entirely disappears: all you can see is your thumb! This is very different from taking at a picture of the whole fist and then looking at the picture close up; the fact that you can still see the whole fist means you're seeing a distortion of reality.

On the other hand, Theorem 3 is not all bad news. If the viewing distance of a picture is say, 40 inches, and you view the picture from a couple of inches too far away, then your viewing distance is $42/40 = 1.05$ times greater than it should be. Theorem 3 says that an object in the picture will be perceived as being 1.05 times too elongated in the direction of the z-axis, an error of only 5%. Thus Theorem 3 gives us a valuable drawing tip:

- Keep the viewing distance fairly large, so that small errors by the viewer in choosing a viewing distance will not cause noticeable distortions.

Of course, the viewing distance should also be "practical and comfortable for the viewer" as we said earlier.

******** SUMMARY OF RESULTS & DRAWING TIPS ********

******** EXERCISES ********

2. Vanishing Points and Vanishing Lines

Our results so far have given us some useful tips for setting up , executing, and viewing perspective drawings and paintings, but the execution part probably seems rather technical to you. There are too many formulas, and too much reliance on coordinates: what we would like to do is grab paper, pencil, and maybe a straightedge, and begin making nice pictures, while keeping the math hidden in the background and concentrating on the art. That's exactly what the concepts of vanishing points and vanishing lines will help us to do. We'll start with vanishing points.

Suppose we have a straight line L in space with parametric equations

$$x = x_0 + at, \qquad y = y_0 + bt, \qquad z = z_0 + ct,$$

where $-\infty < t < \infty$ and x_0, a, y_0, b, z_0, c are constants and $c > 0$. The condition $c > 0$ means that $\lim_{t\to\infty} z(t) = \infty$, so the line gets infinitely far away from the viewer in the positive z-direction as the parameter t approaches infinity. As usual, we consider the viewer to be located at $(0, 0, -d)$, with $d > 0$. Theorem 1 tells us that the perspective image of L (technically, the image of the part of L with positive z-coordinates) is a subset of a straight line. However, the image is not an entire straight line: in particular, it vanishes abruptly at a point called the vanishing point! To see why this should happen, look at Figure 12. A viewer looks at a line L extending infinitely in the positive z-direction. As the eye gazes at more and more distant points on L, the line of sight eventually becomes parallel to L, and the line L seems to disappear, because the two parallel lines (the line of sight and L) cannot intersect. At the instant this happens, the viewer is looking directly at the vanishing point. The vanishing point of a line can be anywhere in the picture plane, depending on the direction of the line.

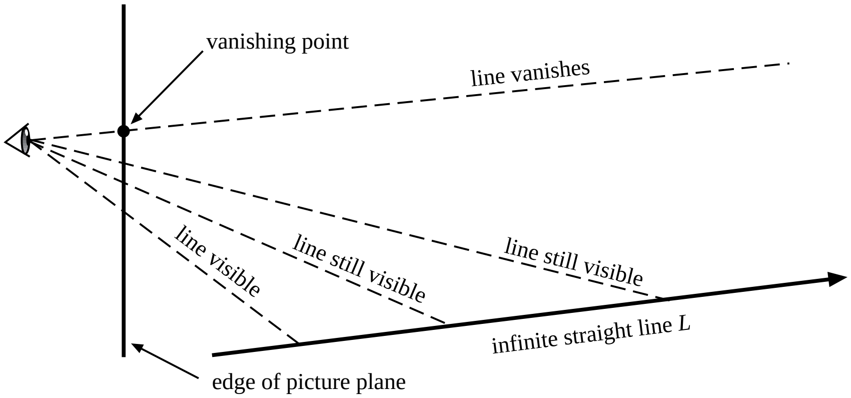


Figure 12.

With the line L given parametrically, it is easy to locate its vanishing point. By (1), the parametric equations of the image of L in the picture plane are

$$x' = \frac{d(x_0 + at)}{(z_0 + ct) + d} = \frac{(da)t + (dx_0)}{(c)t + (z_0 + d)} \qquad \text{and} \qquad y' = \frac{d(y_0 + bt)}{(z_0 + ct) + d} = \frac{(db)t + (dy_0)}{(c)t + (z_0 + d)},$$

where we have arranged terms in the numerators and denominators in descending powers of t. This makes it easy to take limits as t goes to infinity:

$$\lim_{t\to\infty} x' = da/c \qquad \text{and} \qquad \lim_{t\to\infty} y' = db/c.$$

That is, the $x'y'$-coordinates of the vanishing point are $(da/c, db/c)$.

Now we present three results which will help us draw good perspective drawings without explicitly using a lot of technical mathematics. The first result just confirms what we assumed in the above argument, and the other two follow from the first.

Theorem 4. *Let L be a line in space not parallel to the picture plane, with parametric equations*

$$x = x_0 + at, \qquad y = y_0 + bt, \qquad z = z_0 + ct,$$

where $c > 0$. Then the line of sight from the viewpoint $E(0, 0, -d)$ to the vanishing point of L is parallel to L.

Proof. If V denotes the vanishing point of L, then the space coordinates of V are $(da/c, db/c, 0)$, and the line of sight is parallel to $\overrightarrow{EV}$. But

$$\overrightarrow{EV} = \langle da/c, db/c, d\rangle = (d/c)\langle a, b, c\rangle,$$

and $\langle a, b, c\rangle$ is the direction vector of L. ■

Theorem 5. *If two lines L_1 and L_2 are parallel to each other, but not parallel to the picture plane $z = 0$, then they have the same vanishing point.*

Proof. If V_1 and V_2 denote the respective vanishing points (in space coordinates) of L_1 and L_2, then by Theorem 4, $\overrightarrow{EV_1}$ and $\overrightarrow{EV_2}$ must also be parallel. This can only happen if $V_1 = V_2$. ■

Theorem 6. *If a line L in space is parallel to the z-axis, then the vanishing point V of L is the origin $(0, 0)$ of the picture plane $z = 0$.*

Proof. The proof is left to the reader.

It's quite common to combine these two results in a painting or drawing. One example is Figure 13, which features Piero Della Francesca'a painting, *Ideal City*, executed in the 15th century. Like many Renaissance artists, Piero was knowledgeable in mathematics, and wrote treatises on solid geometry and perspective. The perspective setup here is our standard one, with the viewer's line of sight (the z-axis) normal to the picture plane and parallel to the plane of the ground. The principal straight lines of the buildings fall mainly into two categories. There are vertical and horizontal lines in planes parallel to the picture plane: by Theorem 2 their images are also vertical or horizontal, respectively. The other category consists of lines which are parallel to the z-axis, and by Theorems 5 and 6, their images all converge to the origin of the picture plane, as indicated in Figure 14. This situation is so common in art, illustration, and architecture that it has a name. When one point serves as the vanishing point for all, or almost all, of those lines in a painting or drawing which have a vanishing point, it's called "one-point perspective."

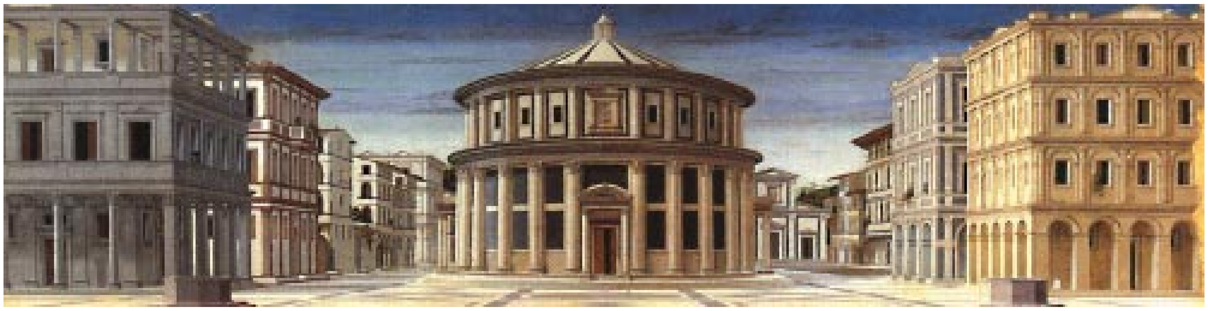


Figure 13. Piero Della Francesca, *Ideal City*

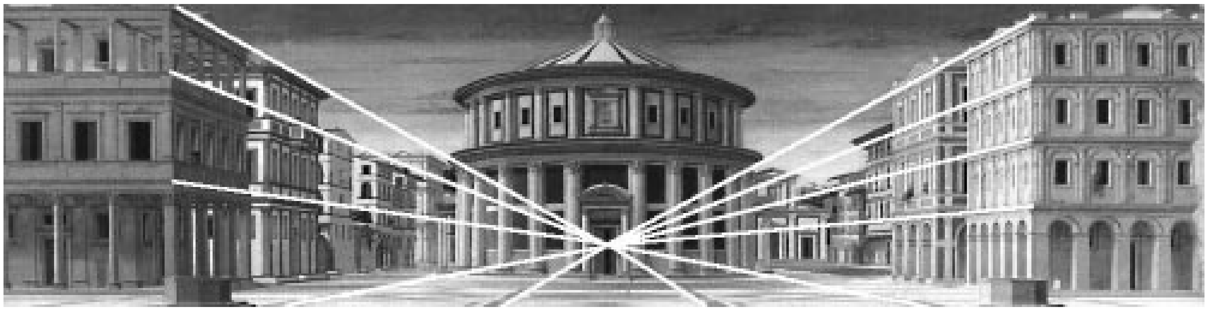


Figure 14. *Ideal City*, with lines converging to a vanishing point.

Of course, a little more was required in this painting. For instance, the two octagonal structures in the foreground have some lines which do not fall into the two categories just mentioned, and the problem of drawing the cylindrical building in the center had to be solved. We'll deal with this type of problem later on.

A famous example of the use of one-point perspective is *The Last Supper* by Leonardo da Vinci (Figures 15 and 16). In this work, the head of Christ is placed near the single vanishing point, so that lines of the architecture parallel to the z-axis (in our terminology) will lead the viewer's eye to the central figure. Thus artists not only use perspective as a means to achieve an illusion of depth, but also as a compositional device to direct the viewer's eye.

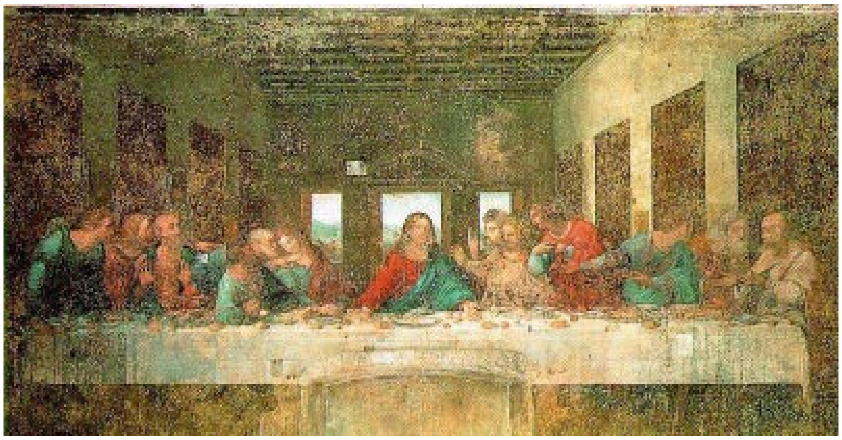


Figure 15. Leonardo da Vinci, *The Last Supper*

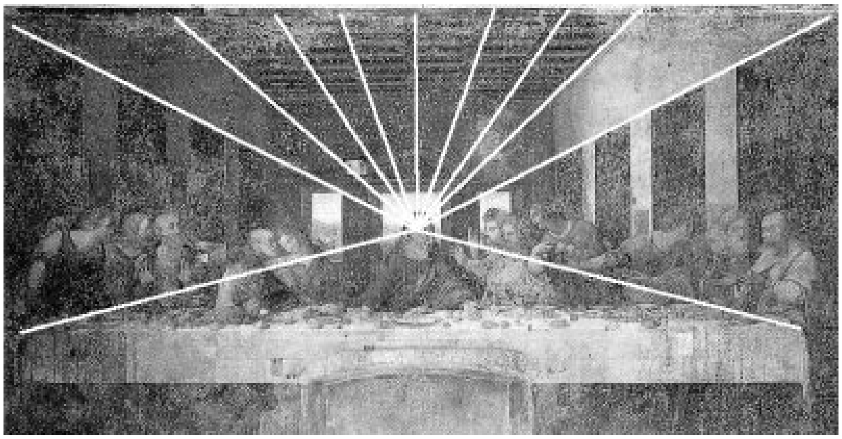


Figure 16. *The Last Supper*, with lines converging to a vanishing point.

Of course, one vanishing point is often not enough to represent scenes realistically. The building in the photograph in Figure 17, when viewed as shown, requires two vanishing points for an adequate representation. When just two vanishing points are used, the siuation is often referred to as "two-point perspective."

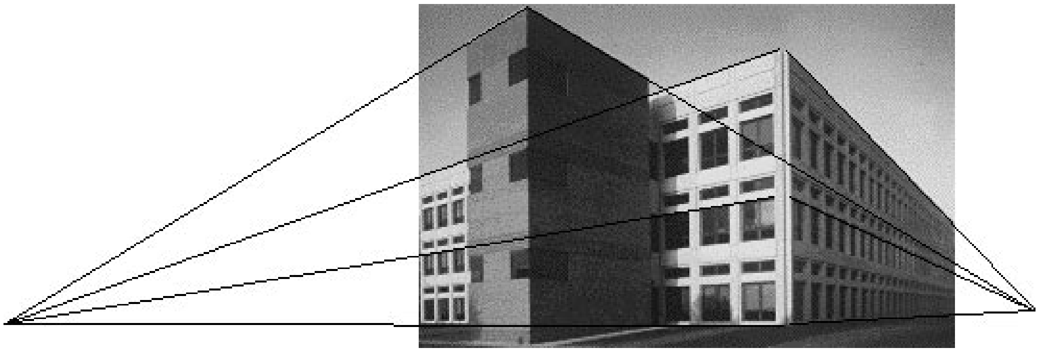


Figure 17. A scene with two vanishing points.

Notice that lines of the building which are vertical in space have images that are also vertical—and more importantly, parallel—in the picture. This indicates that the picture plane is parallel to these lines; otherwise, by Theorem 4, they would converge to a vanishing point.

An example of just such a situation is the photograph in Figure 18. The photo was taken looking upward at a skyscraper, and hence the picture plane is no longer parallel to the vertical lines of the building. As a result, the images of these lines converge to a vanishing point. There are three principal vanishing points in this image—a fairly common occurence in architectural renderings. Naturally, this situation is referred to as "three-point perspective."

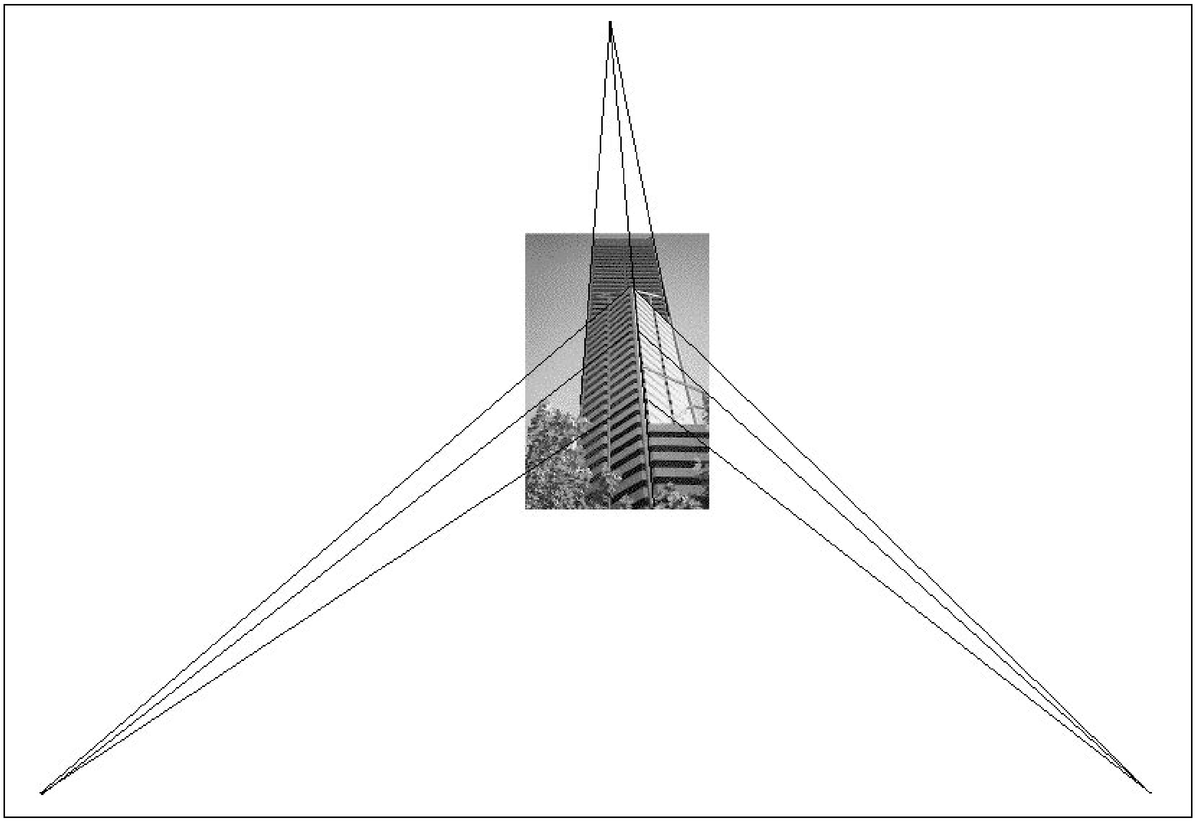


Figure 18. A scene with three vanishing points.

In practice, a perspective drawing may have any number of vanishing points, depending on what kind of scene one is required to draw. A concept which is often helpful in locating vanishing points is that of the vanishing line of a plane. To visualize this, look at the aerial photograph in Figure 19. As shown in Exercise #, the part of the earth we can see locally can be idealized as a plane, provided that the terrain is not too hilly and we are not at too high an altitude.

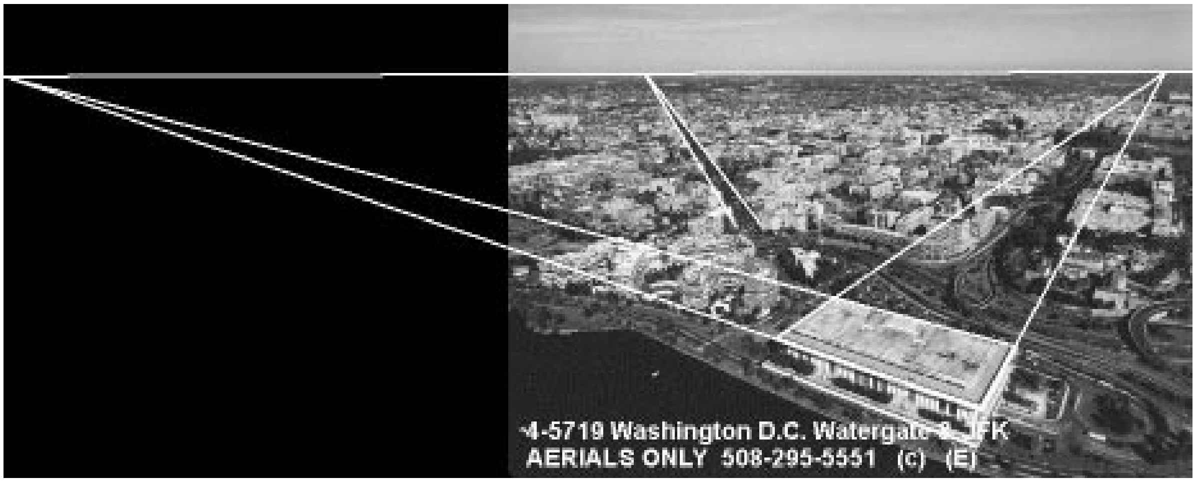

4-5719 Washington D.C. Watergate & JFK
AERIALS ONLY 508-295-5551 (c) (E)

Figure 19. Three vanishing points on a vanishing line.

Now notice that there are three pairs of of lines in the picture, each pair converging to a different vanishing point. The corresponding lines in the real world are architectural lines of buildings or streets, and the two members of any pair are in actuality parallel to one another, but not parallel to the picture plane; hence the convergence of the images of any given pair to a single vanishing point is guaranteed by Theorem 5. The interesting thing here is that all three vanishing points lie on the same line, which coincides with the horizon in the picture. The reason this happens is that all the lines are parallel to the plane of the ground. The horizon line in this instance is an example of what we call a vanishing line. One way of looking at a vanishing line is to consider it a set of vanishing points for other lines, such as the lines of streets and buildings we just mentioned. Another way of looking at the vanishing line is to think of it as the place where an entire plane seems to vanish: It turns out that if the earth were actually an infinite flat plane, it would not appear that different to us at relatively low altitudes, and in fact, even though it would no longer be curved, it would still appear to vanish at almost the exact location of the horizon line in the picture.

To make some of these ideas more precise—and more useful to our drawing needs—we present the next two results on vanishing lines.

Theorem 7. *Let Γ be a plane in space given by the equation*

$$Mx + Ny + Pz + Q = 0,$$

where M and N are not both 0; i.e., Γ is not parallel to the picture plane $z = 0$. If the viewer is located as usual at $(0, 0, -d)$, then any line L in Γ which is not parallel to the picture plane has a vanishing point which lies on the line

$$Mx' + Ny' = -dP,$$

called the vanishing line *of the plane Γ.*

Proof. Let L be a line in Γ not parallel to the picture plane, and let us consider a point (x, y, z) on L. Assume that L is parameterized as $\langle x(t), y(t), z(t)\rangle$, where $z \to \infty$ as $t \to \infty$, and let (x'_v, y'_v) be the vanishing point of L, so that $x'_v = \lim_{t\to\infty} x'$ and $y'_v = \lim_{t\to\infty} y'$. Then

$$\begin{aligned} Mx'_x + Ny'_v &= M \lim_{t\to\infty} x' + N \lim_{t\to\infty} y' = \lim_{t\to\infty} (Mx' + Ny') \\ &= \lim_{t\to\infty} \left(M\frac{dx}{d+z} + N\frac{dy}{d+z} \right) = \lim_{t\to\infty} \frac{d(Mx+Ny)}{d+z}. \end{aligned}$$

But L lies in Γ, so for any value of t, the point (x, y, z) does also. Hence $(Mx + Ny) = -Pz - Q$, and we can write

$$Mx'_x + Ny'_v = \lim_{t\to\infty} \frac{-dPz - dQ}{z + d} = \lim_{z\to\infty} \frac{-dPz - dQ}{z + d} = -dP.$$

It follows that (x'_v, y'_v) lies on the vanishing line. ■

Corollary 8. *If Γ_1 and Γ_2 are parallel planes, not parallel to the picture plane $z = 0$, then they have the same vanishing line.*

Proof. If the equation of Γ_1 is put in the form $Mx + Ny + Pz + Q_1 = 0$, then Γ_1 has a normal vector $\langle M, N, P\rangle$. Since Γ_2 is parallel to Γ_1, it has the same normal vector, so its equation can be put in the form $Mx + Ny + Pz + Q_2 = 0$ for some number Q_2. But Theorem 7 says that the equation of the vanishing line depends only on M, N, P, and the viewing distance d; that is, the two planes have the same vanishing line $Mx' + Ny' = -dP$. ■

Figure 20 illustrates a common application of Theorem 7. To get the main idea, let's not worry about viewing distance or object size. Our goal is to draw a structure with a slanted roof, like a house, a children's playhouse, or a tool shed. The steps are as follows.

(a) With our line of sight level with respect to the ground, the plane of the ground is parallel to the plane $y = 0$. Thus the equation of the ground plane can be put in the form $Mx + Ny + Pz + Q = 0$, with $M = P = 0$ and $N = 1$. By Theorem 7, the vanishing line of the ground plane (commonly called the horizon line) is the line $y' = 0$ (the x'-axis). Thus we choose a horizontal line as the horizon, and shade the ground gray to make it stand out.

(b) Let's make the front wall of the building parallel to the picture plane $z = 0$, so that it is an undistorted rectangle (a consequence of Theorem 2). Let's also make the left wall visible. Since the left wall is perpendicular to the front wall, its top and bottom edges, which are parallel to the ground plane, are also parallel to the z-axis. Thus by Theorem 6, the vanishing point v_0 for these lines is the origin $(0, 0)$ in the picture plane. This vanishing point must therefore be on the horizon line $y' = 0$ and somewhere to the left of the corner of the building, so that we can see the left wall. Since we're not concerned about object size or viewing distance, we place v_0 on the horizon somewhat arbitrarily, and use it to draw the left wall, whose back edge we also locate somewhat arbitrarily. To repeat, v_0 is now the origin of the picture plane, and thus our center of view is well to the left of the center of the picture in this exercise.

(c) In order to locate the peak of the roof, we first need to locate the center of the left wall. Since the diagonals of the left wall cross at the center of the wall, and since their perspective images are also line segments, we can locate the perspective image of the center of the wall by simply drawing the images of the diagonals (dashed lines) and noting their intersection point. From this point we draw a vertical line segment to locate the peak of the roof at an arbitrary height. The question now is: How do we correctly draw the rest of the roof?

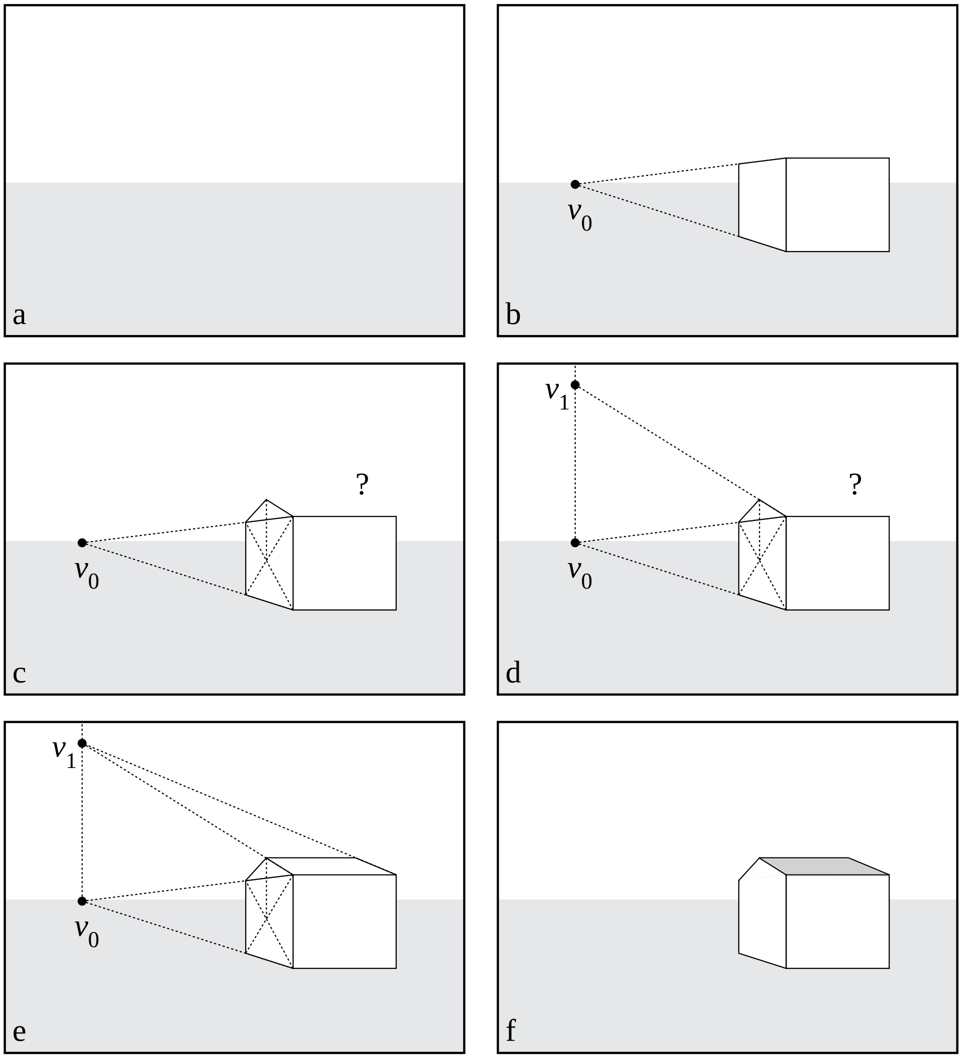


Figure 20. An application of Theorem 7.

(d) The answer is given by Theorem 7. The plane of the left wall has an equation of the form $Mx + Ny + Pz + Q = 0$, where $N = P = 0$. (Why?) Thus by the Theorem 7, the vanishing line of this plane is the line $x' = 0$ (the y'-axis). In particular, the left front edge of the roof has its vanishing point v_1 on this line, so we locate it by simply extending the line of the roof edge until it meets the vertical line $x' = 0$.

(e) The right front edge of the roof is parallel to this line, so by Theorem 5 it has the same vanishing point. Hence we can draw the right front edge, and then draw the top edge, which must be horizontal since it is level with the ground and parallel to the picture plane.

(f) Finally, we erase all the construction lines, add a little color to the roof, and we are done.

Our building seems a little lonely, so let's add a road with a fence on either side of it

(Figure 21). To do this, we can again employ a step-by-step method which makes use of Theorem 7.

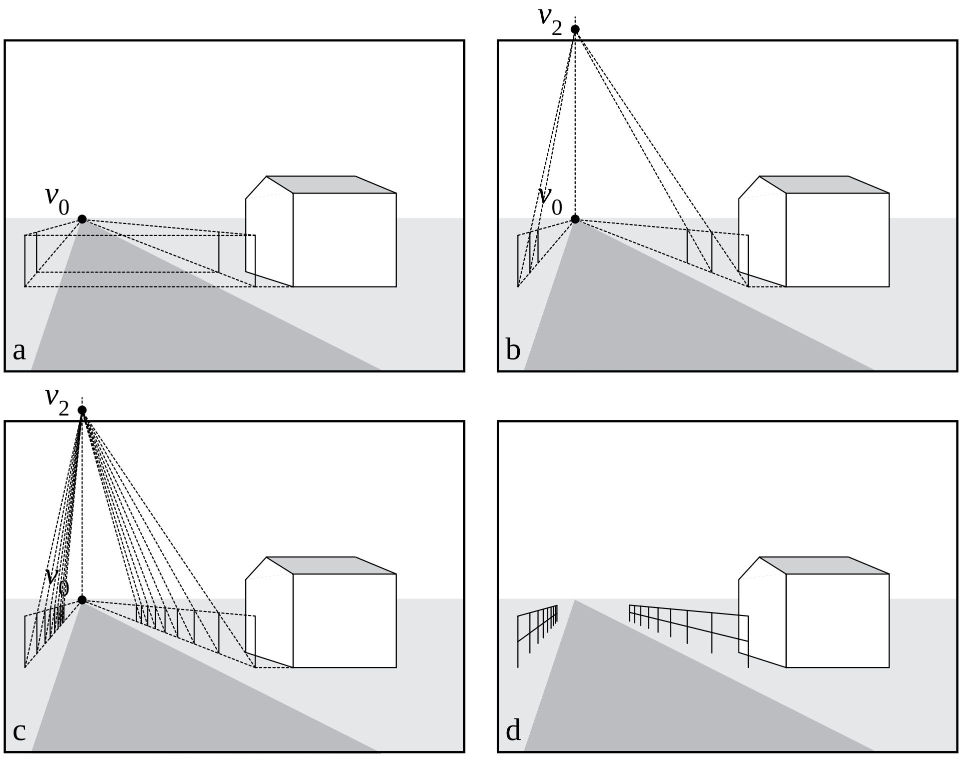


Figure 21. Using Theorem 7 to locate fenceposts.

(a) If we choose the road to be parallel to the left wall of the building, we can use our previous vanishing point v_0 (the origin of the picture plane). After shading the road, we can locate the first fenceposts even with the front corner of the building by sketching a dashed horizontal line from the corner of the building and making the posts have equal height. The fences are parallel to the road, so their top and bottom edges have v_0 as their vanishing point, as indicated by the dashed lines. We locate the second post on the right somewhat arbitrarily, and make the second post on the left even with it by use of another horizontal dashed line. We can't be arbitrary about the rest of the fenceposts, however, because fenceposts have equal spacing.

(b) In reality, the rectangular sections of fence between the fenceposts all have the same shape, so their diagonals are parallel; thus the perspective images of these diagonals will all have the same vanishing point v_2. To locate v_2, note that the plane of each fence is parallel to the plane of the left wall of the building, so by Corollary 8 they have the same vanishing line, which was previously determined to be the y'-axis. We re-draw the y'-axis through v_0 and extend the diagonal from the bottom of the first fencepost through the top of the second fencepost until it intersects the y'-axis; this locates v_2. Next, we draw the dashed line from the bottom of the second fencepost to v_2, and the place where this line intersects the top edge of the fence determines the top of the third fencepost. We the draw the third fencepost on either side.

(c) Continuing this process, we keep going until we're tired of drawing fenceposts. Actually, we shouldn't go too far unless we're going to draw our lines thinner and thinner as we go into the distance.

(d) Finally, we erase our construction lines and add another rail to each fence, using the vanishing point v_0 at the "end" of the road.

The method we just used to locate fenceposts isn't the only one, or even the best one. For instance, the vanishing point v_2 was off the "canvas," and this is sometimes inconvenient. Another method is presented in Exercise #.

******** EXERCISES ********

3. Analyzing Works Done in Perspective

The theorems we have presented so far are not just good for creating works in perspective. We can also use them to help us better understand and appreciate paintings and drawings that other people have done in perspective. To begin with, let's practice on the building we drew in Figure 20. We said we wouldn't worry about viewing distance or object size, but now that the drawing is done, is there anything that can be said about these things? The answer is yes. To prove it, we're going to play dumb and forget the information we used to set up the drawing, and see what can be deduced just by looking. As each deduction is stated, try to flesh out the argument or name the theorem which justifies it.

Let's start with the viewing distance. We saw in Theorem 3 that the viewing distance is related to the shape of the object being drawn. If we consider our building to be, say, a backyard toolshed, then we can make some reasonable assumptions about its shape and size. To analyze the picture, we first notice that the front wall is apparently a perfect rectangle (Figure 22(a)), so it is parallel to the picture plane (otherwise, at least one pair of its edges would converge to a vanishing point instead of being parallel). Next, we extend the lines of the top and bottom edges of the left wall to locate the vanishing point v_0. Since the left wall is in reality orthogonal to the front wall, its top and bottom edges are parallel to our line of sight (the z-axis), so the vanishing point v_0 must be the origin of the picture plane. Thus, to view the picture correctly, we should place ourselves (rather, one eye) directly in front of v_0, but how far from the picture? To find out, we draw the y'-axis, which is the vanishing plane of the left wall. We could make use of the left edge of the roof, but to show that our technique doesn't depend on slanted roofs, we instead extend a diagonal of the rectangular part of the left wall until it meets the y'-axis at a new vanishing point v_3, and we label the distance from v_0 to v_3 with the letter b.

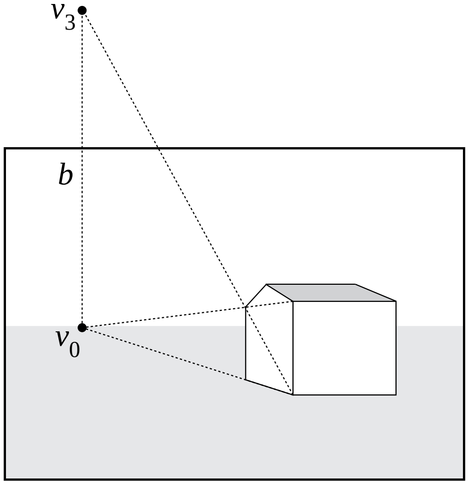


Figure 22(a).

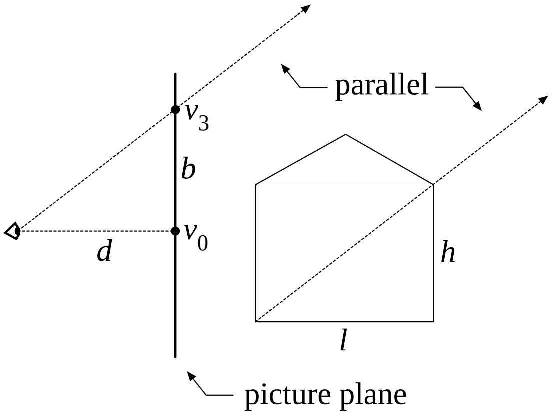


Figure 22(b).

Figure 22(b) shows a side-view schematic of the situation. In space, the diagonal and the line from the viewer's eye through v_3 must be parallel by Theorem 4, and hence by similar triangles, we have $d/b = l/h$, where d is the viewing distance and l and h are the actual length and height, respectively, of the wall. Thus we have

$$d = b(l/h).$$

Notice that we don't need to know l and h to compute d, but only their "shape ratio" l/h. If we consider Figure 22(b) as a reasonable sketch of the wall of a toolshed, then $l/h \approx 1.3$. This means that the correct viewing distance is only about 1.3 times as long as b, so our somewhat haphazard drawing has resulted in a viewing distance which is uncomfortably small.

Now try this technique out for yourself by estimating the viewing distance for Figure 13—you'll need a little cleverness!

Determining the viewing distance is an exercise which must be done on a case-by-case basis, and sometimes it's not possible to do with much accuracy. For instance, a photograph of a cloud would be difficult to analyze without more sophisticated knowledge of meterology, the specific camera used, etc. As we have seen, the presence of architecture helps a lot. As an example of a somewhat more difficult problem involving buildings, consider the photograph in Figure 23.

This figure is an aerial photograph of Franklin & Marshall College in Lancaster, Pennsylvania. The image was taken from the F&M home page at `http://www.FandM.edu/`. The first problem we must face is the fact that it is not as easy to be sure of the center of view—that is, the standard origin of the $x'y'$-coordinate system—in the picture plane. A good guess would be the center of the photograph, which would be true if the photo has not been cropped. However, we will not make this assumption, and we will instead proceed as follows.

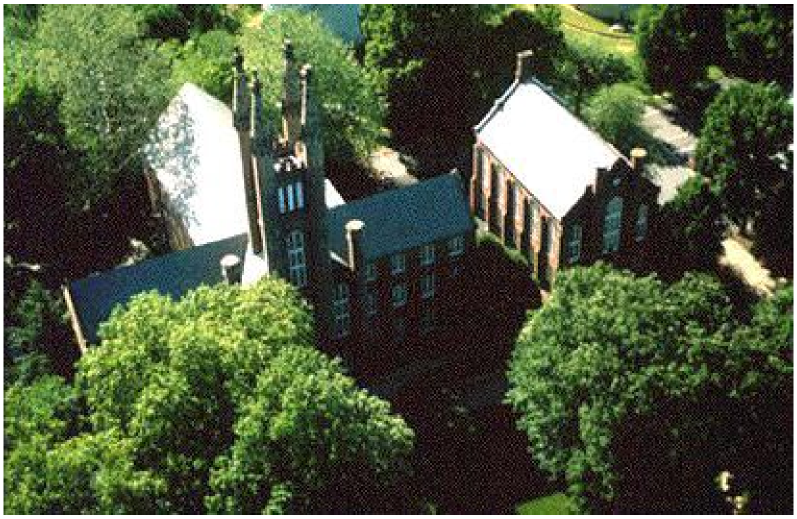


Figure 23. Franklin & Marshall College.

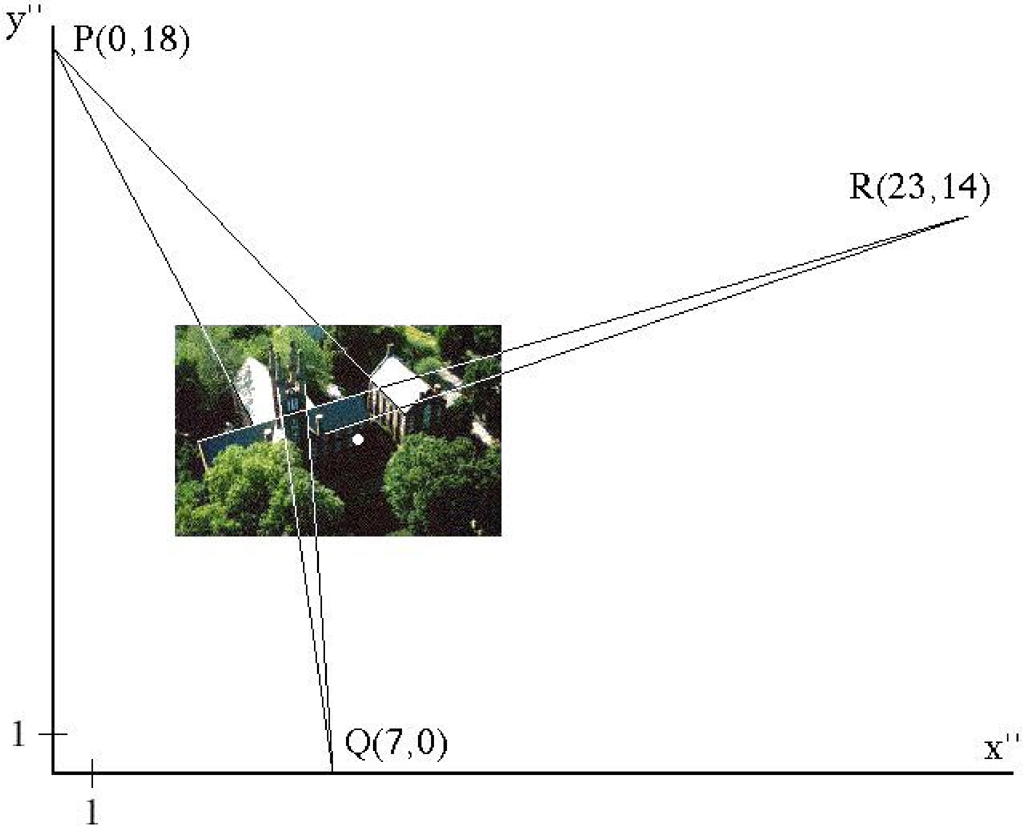


Figure 24. Locating vanishing points in the picture plane.

First, as illustrated in Figure 24, we put the photo into the picture plane (which can be a large piece of paper) and extend pairs of parallel roof lines and wall lines until they meet at the vanishing points P, Q, and R. Next, we mark off a pair of perpendicular axes parallel to the edges of the photo, and choose them for convenience to contain two of the vanishing points. Since these axes are almost certainly *not* the actual $x'y'$-axes, we label them x'' and y''. We then choose a unit of length, indicated by the 1's on the axes, and estimate the $x''y''$-coordinates of P, Q, and R. In order to proceed further, we need to visualize the viewer and the picture plane in 3-dimensional space.

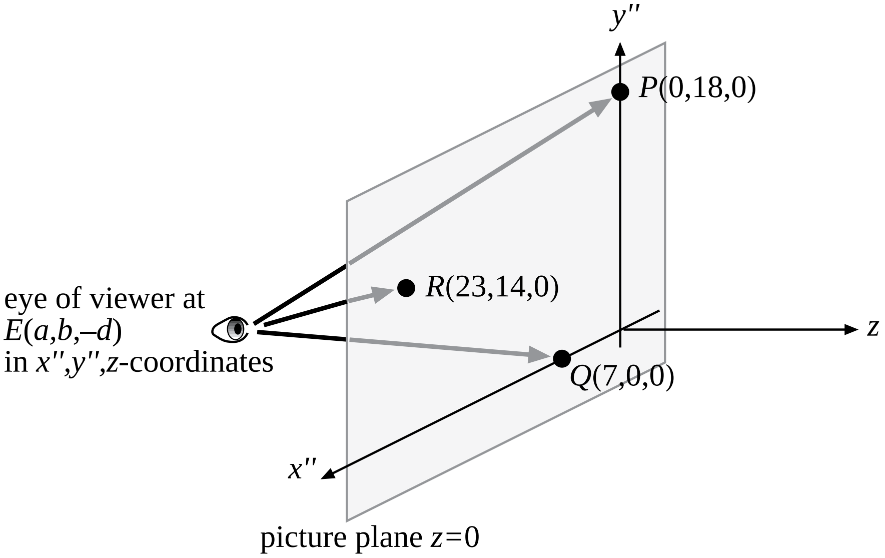


Figure 25. The points E, P, Q, and R in $x''y''z$-coordinates.

In Figure 25 we make use of the standard z-axis, with the picture plane being $z = 0$ as usual. However, since we only have $x''y''$-coordinates available at this stage, we use x'', y'', and z as our space coordinates. Thus the viewer's eye is located at the point $E(a, b, -d)$ in $x''y''z$-coordinates, and it is our job to determine a, b, and d. By Theorem 4, the vectors $\overrightarrow{EP}$, $\overrightarrow{EQ}$, and $\overrightarrow{ER}$ must be parallel to the wall and roof lines whose images converge respectively to P, Q, and R in the picture plane. But in space, these architectural lines are either parallel or perpendicular to each other, and thus $\overrightarrow{EP}$, $\overrightarrow{EQ}$, and $\overrightarrow{ER}$ are mutually orthogonal. Hence we can write

$$\overrightarrow{EP} \cdot \overrightarrow{EQ} = 0, \qquad \overrightarrow{EP} \cdot \overrightarrow{ER} = 0, \qquad \overrightarrow{EQ} \cdot \overrightarrow{ER} = 0,$$

or, respectively,

$$\begin{aligned} \langle -a, 18 - b, d\rangle \cdot \langle 7 - a, -b, d\rangle &= 0 \\ \langle -a, 18 - b, d\rangle \cdot \langle 23 - a, 14 - b, d\rangle &= 0 \\ \langle 7 - a, -b, d\rangle \cdot \langle 23 - a, 14 - b, d\rangle &= 0. \end{aligned}$$

Carrying out the dot products, we get, respectively,

$$\begin{aligned}
a^2 - 7a + b^2 - 18b + d^2 &= 0 \\
a^2 - 23a + b^2 - 32b + 252 + d^2 &= 0 \\
a^2 - 30a + 161 + b^2 - 14b + d^2 &= 0.
\end{aligned}$$

If we now subtract the second equation from the first, and then subtract the third equation from the first, we get the simple system

$$\begin{aligned}
16a + 14b &= 252 \\
23a - 4b &= 161
\end{aligned}$$

whose solution is

$$a \approx 8.45, \qquad b \approx 8.34.$$

Finally, we can substitute these values into any of the equations involving d to get

$$d \approx 8.3.$$

Actually, our work is not that accurate, so we would be more honest to write

$$a \approx 8, \quad b \approx 8, \quad d \approx 8,$$

and conclude that the viewing distance is about 8 units for the small version of the photo. The center of view (the true $x'y'$ origin) has been indicated in Figure 24 as a white dot at the point $(8.45, 8.34)$ in $x''y''$-coordinates. The fact that it is off-center in the photo is most likely an artifact our own experimental errors in drawing lines and estimating coordinates, rather than a result of the photo being cropped.

If all these computations seem tedious, watch how easy it is in the computer algebra program Maple. First, we load the linear algebra package, which handles vectors; at the next prompt, we define the vectors $\overrightarrow{EP}$, $\overrightarrow{EQ}$, and $\overrightarrow{ER}$; finally, we solve the dot product equations for a, b, and d, using the fsolve function (the "f" is for "floating-point" as in decimal point).

```
> with(linalg):
> EP:=[0,18,0]-[a,b,-d]:  EQ:=[7,0,0]-[a,b,-d]:  ER:=[23,14,0]-[a,b,-d]:
> fsolve({dotprod(EP,EQ)=0, dotprod(EP,ER)=0, dotprod(EQ,ER)=0}, {a,b,d});
            {a = 8.450777202, b = 8.341968912, d = 8.264792809}
```

WARNING: The results obtained here are very rough approximations. When you start doing this type of analysis yourself, you'll see that no two people exactly agree on how to extend roof or wall lines to vanishing points. If you do it twice, you may not even agree with yourself! When the images of roof or wall lines are very nearly parallel—and they frequently are—slight changes in the lines you draw to extend them can make big differences in the estimated location of vanishing points, and correspondingly large differences in the computed viewing distance. We have shown that the viewing distance is computable in *principle*, but in practice, it can be tricky.

Find the center of view (the $x'y'$ origin) and compute the viewing distance for Figure 18.

So far we have seen some examples of how to compute the correct distance from the viewer to the picture plane, but how about the distance from the viewer to the objects in the picture? Of course, the actual objects could be miles away or long vanished; they may never have even existed, like the little shed we drew in Figure 20. Nevertheless, the question can be meaningful in certain cases. For instance, whenever a work in perspective is based on a real scene, it is reasonable to expect that the artist viewed the canvas from what we would call the "correct" point of view, and if the scene appeared to the artist as it does in the painting from that point, then we can attempt to "reverse engineer" the picture and estimate how far from the objects in the picture the artist was located. In the case of photographs, we can try to determine where the camera was located in the scene. In some cases, the estimate can be made rather easily, but first we'll attempt the analysis on the photograph in Figure 23 by placing it in the picture plane again (Figure 26).

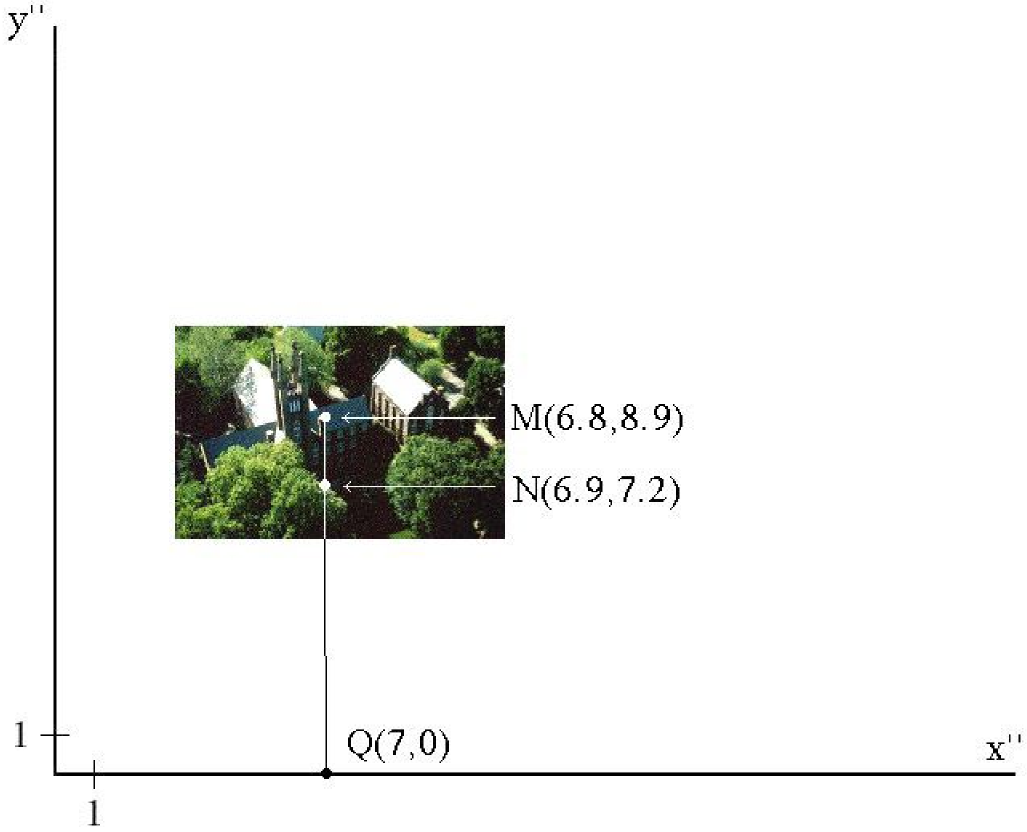


Figure 26.

The photograph was apparently taken from a considerable height, so we can ask, How far above the ground was the camera? At first it may seem that we couldn't say anything about that; perhaps the photo was taken from a spy plane or satellite, using high-resolution photography. But in fact—although we can't make a highly accurate determination, as we indicated above—we can use our knowledge of perspective and some vector mathematics to rule out any covert spying on Franklin & Marshall College! The first step is to pick an object in the picture whose actual size we can reasonably estimate; we choose the cylindrical turret to the right of the main entrance, and we label the images of points on the top and bottom as M and N, respectively, in Figure 26. The coordinate system is the same as that in Figure 24. The coordinates of M and N were actually estimated on a higher resolution image, so we dare to name them to one decimal place. The point Q is the same as in Figure 24; i.e., it's the vanishing point of the images of all the vertical architectural lines, so it's collinear with M and N. Now, as we did earlier, we pass from $x''y''$-coordinates to $x''y''z$-coordinates and view the whole setup in 3-dimensional space (Figure 27.)

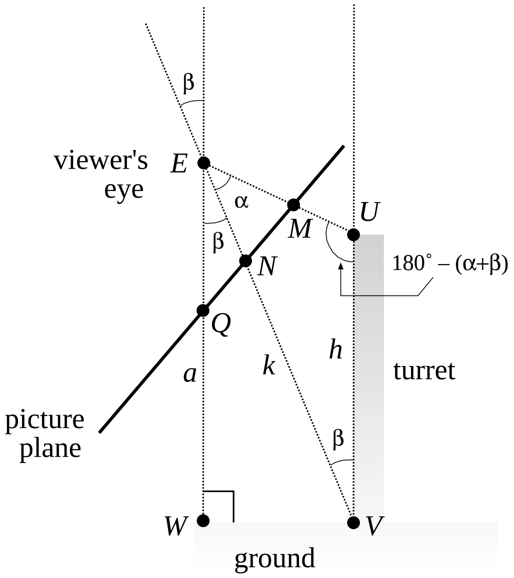


Figure 27.

We have drawn the picture plane (seen edge-on) as being tilted, in order to have the turret appear vertical. Since we still consider the picture plane to be the plane $z = 0$, the whole coordinate system is therefore tilted in the picture. The known quantities are the height h of the turret (which we will estimate later) and the points E, M, N, and Q. If we round our previously estimated $x''y''z$-coordinates of E to one decimal place, we have

$$E = (8.5, 8.3, -8.3), \quad M = (6.8, 8.9, 0), \quad N = (6.9, 7.2, 0), \quad \text{and} \quad Q = (7, 0, 0).$$

We denote by k the distance from E to V. The line from E through the vanishing point Q of all vertical lines must itself be vertical, and hence perpendicular to the ground at W. Thus the altitude a of the viewpoint E—the quantity we are after—is the distance from

E to W. Finally, you should convince yourself that our labeling of all the angles makes sense.

Applying the law of sines to triangle EUV, and using the fact that $\sin(180° - (\alpha + \beta)) = \sin(\alpha + \beta)$, we have

$$\frac{k}{\sin(180° - (\alpha + \beta))} = \frac{h}{\sin \alpha} \implies \frac{k}{\sin(\alpha + \beta)} = \frac{h}{\sin \alpha} \implies k = h \cdot \frac{1}{\sin \alpha} \cdot \sin(\alpha + \beta).$$

Using the right triangle EVW, we can now find a:

$$a = k \cos \beta = h \cdot \frac{1}{\sin \alpha} \cdot \sin(\alpha + \beta) \cdot \cos \beta,$$

and the last three factors can all be written in vector notation:

$$\begin{aligned} a &= h \cdot \frac{\|\overrightarrow{EM}\| \|\overrightarrow{EN}\|}{\|\overrightarrow{EM} \times \overrightarrow{EN}\|} \cdot \frac{\|\overrightarrow{EM} \times \overrightarrow{EQ}\|}{\|\overrightarrow{EM}\| \|\overrightarrow{EQ}\|} \cdot \frac{\overrightarrow{EN} \cdot \overrightarrow{EQ}}{\|\overrightarrow{EN}\| \|\overrightarrow{EQ}\|} \\ &= h \frac{\|\overrightarrow{EM} \times \overrightarrow{EQ}\| (\overrightarrow{EN} \cdot \overrightarrow{EQ})}{\|\overrightarrow{EM} \times \overrightarrow{EN}\| \|\overrightarrow{EQ}\|^2}. \end{aligned}$$

Since $\|\overrightarrow{EQ}\|^2 = \overrightarrow{EQ} \cdot \overrightarrow{EQ}$, it might be prettier to write a as

$$a = h \frac{\|\overrightarrow{EM} \times \overrightarrow{EQ}\| (\overrightarrow{EQ} \cdot \overrightarrow{EN})}{\|\overrightarrow{EM} \times \overrightarrow{EN}\| (\overrightarrow{EQ} \cdot \overrightarrow{EQ})}.$$

If we estimate the height h of the turret at about 60 feet, we can compute a in Maple:

```
> with(linalg):
> E:=[8.5,8.3,-8.3]:  M:=[6.8,8.9,0]:  N:=[6.9,7.2,0]:  Q:=[7,0,0]:
> EM:=M-E: EN:=N-E: EQ:=Q-E:
> a:=60*norm(crossprod(EM,EQ),2)*dotprod(EQ,EN)/
  (norm(crossprod(EM,EN),2)*dotprod(EQ,EQ));
                          a := 180.2357150
```

Thus the altitude a is about three times the value of h, or about 180 feet. However, since our original measurments of the picture are very rough approximations, we should be cautious and simply say that the camera was no higher than a few hundred feet off the ground.

As we said earlier, estimating the distance to actual objects whose images appear in a picture is not always so hard. For instance, the doorway to the cylindrical building in Figure 13 is in a plane parallel to the picture plane (How do we know?) and its image contains the central vanishing point, i.e., the center of view. Assuming you already computed the viewing distance d, it is now easy to find the distance c from the viewer to the doorway. A side view of the situation appears in Figure 28, where we have denoted the points on the bottom and top of the image of the doorway directly below and above the vanishing point by A' and B', respectively, and the actual bottom and top of the doorway are denoted respectively by A and B. The height of the image of the doorway is h' and the actual height of the doorway is denoted by h.

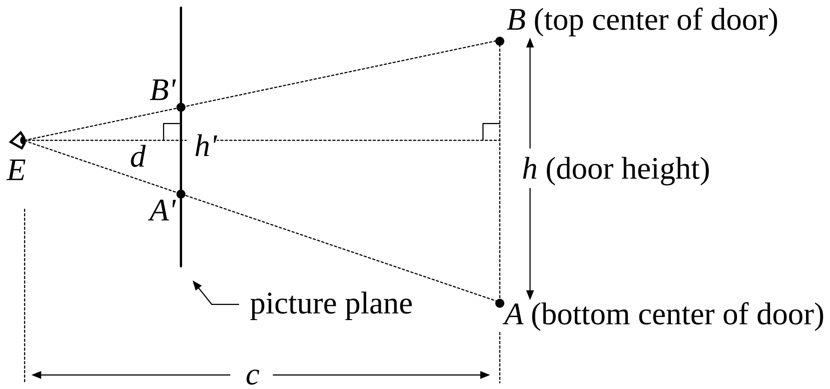


Figure 28.

If we let E denote the location of the viewer, then triangles $EA'B'$ and EAB are similar, and hence the ratios of corresponding altitudes to corresponding sides are equal; that is, $c/h = d/h'$. It follows that

$$c = \frac{dh}{h'},$$

or, in words,

$$\text{(distance from viewer to door)} = \frac{\text{(viewer-to-image distance)(actual door height)}}{\text{(height of door image)}}.$$

If you express d and h' in inches, convert to feet, and then estimate h in feet, you should get a reasonable estimate for c in feet. Try it!

Another case in which the viewing distance is not hard to compute is the case of two-point perspective, such as in Figure 17 or Figure 29 below. By "two-point perspective" we mean that a picture has been set up in such a way that the picture plane is perpendicular to the plane of the ground or floor, and only two vanishing points on the horizon line are needed to render buildings or other objects whose adjacent vertical walls or sides are perpenducular to each other. This is a typical situation when dealing with architectural subjects, both indoors and outdoors.

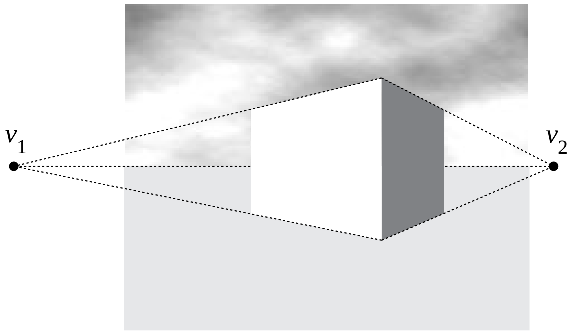


Figure 29.

Suppose the rectangular box in Figure 29 represents some kind of building, but we don't know anything about its size or proportions. Can we say anything about the correct location of the viewer? It turns out that we can. First, notice that we could guess by looking that the picture plane is perpendicular to the plane of the ground—that is, vertical—because the images of the vertical lines of the building are parallel to one another in the picture, and hence do not converge to a vanishing point. This means that the actual vertical lines of the building are parallel to the picture plane, so the picture plane is perpendicular to the plane of the ground. From the discussion in Step (a) following Corollary 8, this means that the horizon line is also the x-axis, so the center of view lies somewhere on this line. Moreover, since our line of sight is level, the viewer location is somewhere in a horizontal plane containing the horizon line of the picture. This horizontal plane is of course the plane $y = 0$. This situation is typical of two-point perspective drawings. The question is, where in the plane $y = 0$ should the viewer be located? (See Figure 30).

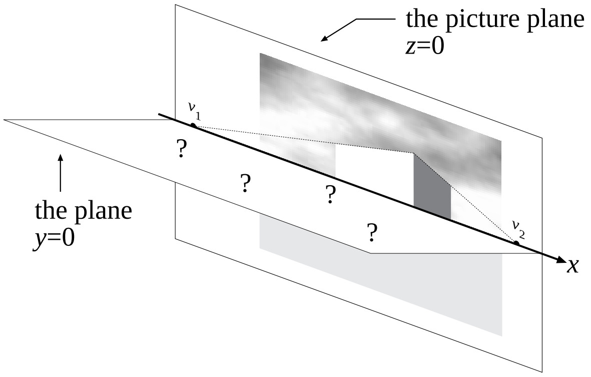


Figure 30.

Since the viewer location is unknown—as indicated by the question marks—we don't yet know where to put the y-axis or the z-axis. In order to use coordinates to analyze the problem, let us choose the midpoint of the two vanishing points v_1 and v_2 in Figure 30 as an origin, and introduce a z''-axis through this origin and orthogonal to the picture plane, and an x''-axis which coincides with the x-axis.. This is indicated in Figure 31, where, in each of the two cases shown there, we are looking down onto the plane $y = 0$.

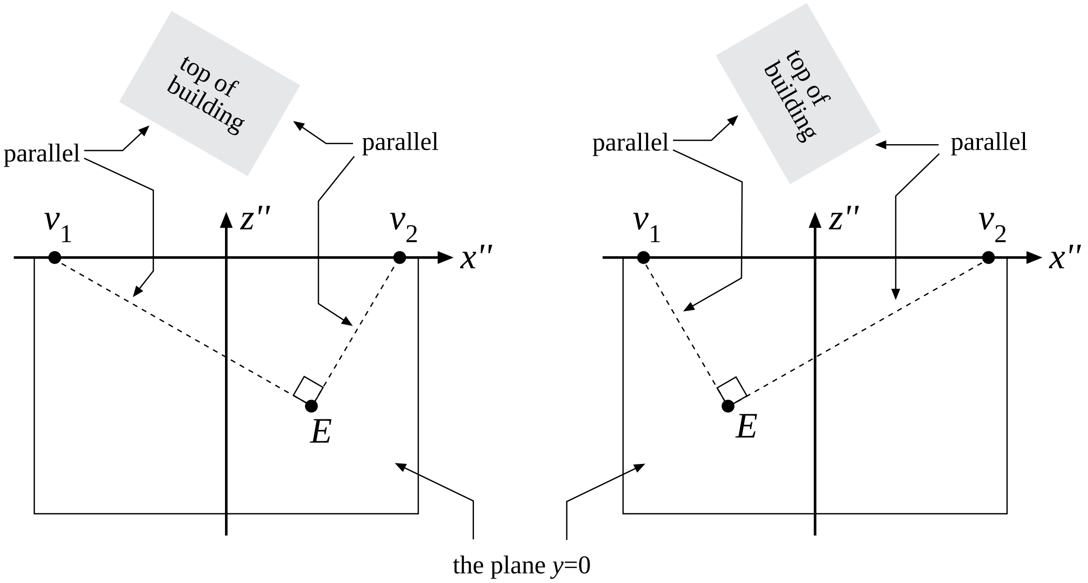


Figure 31. Two possible cases for the viewer location E.

Figure 31 illustrates the fact that, however the viewer and the building are situated, we can be sure of two things:

- The lines of sight from E to the vanishing points v_1 and v_2 are parallel to the horizontal edges of the building.
- These lines of sight must be perpendicular at E, because adjacent horizontal edges of the building are perpendicular.

This really narrows down the possibilities for E. To explain why, we will refer to Figure 32. In the figure we are again looking down onto the plane $y = 0$. We have labeled the vector from the origin to v_2 as $\mathbf{r}$, and since we located the origin at the midpoint of v_1 and v_2, the vector from the origin to v_1 can be labeled $-\mathbf{r}$. The vector from the origin to the viewer location E is labeled $\mathbf{a}$, and you should convince yourself that the vectors $\mathbf{r} - \mathbf{a}$ and $-\mathbf{r} - \mathbf{a}$ are labeled correctly.

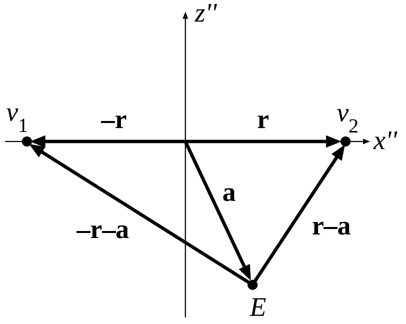


Figure 32.

Now we just observed that $-\mathbf{r}-\mathbf{a}$ and $\mathbf{r}-\mathbf{a}$ must be orthogonal, so we can write

$$(-\mathbf{r}-\mathbf{a})\cdot(\mathbf{r}-\mathbf{a})=0.$$

expanding the left side of this equation, we get

$$-\mathbf{r}\cdot\mathbf{r}+\mathbf{r}\cdot\mathbf{a}-\mathbf{a}\cdot\mathbf{r}+\mathbf{a}\cdot\mathbf{a}=0.$$

Since $\mathbf{a}\cdot\mathbf{r}=\mathbf{r}\cdot\mathbf{a}$, and since $\mathbf{r}\cdot\mathbf{r}=||\mathbf{r}||^2$ and $\mathbf{a}\cdot\mathbf{a}=||\mathbf{a}||^2$, this simplifies to

$$-||\mathbf{r}||^2+||\mathbf{a}||^2=0,$$

and hence

$$||\mathbf{a}||=||\mathbf{r}||.$$

Since $||\mathbf{a}||$ is the distance from the viewer to the midpoint of v_1 and v_2, this means that the correct viewer location lies somewhere on a semicircle of radius $||\mathbf{r}||$ centered at the midpoint of the vanishing points. This fact is interesting enough (and as we'll see later, useful enough) to summarize with a picture and a theorem.

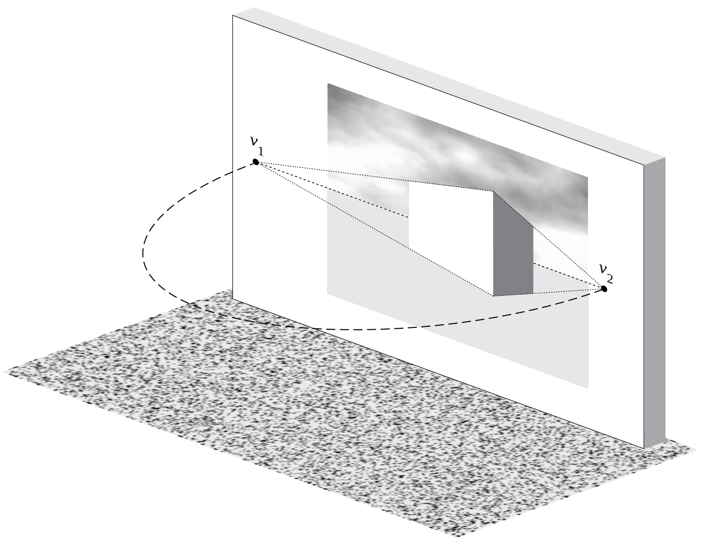


Figure 32. The viewpoint for two-point perspective lies on the semicircle.

Theorem 9. *When viewing a work done in two-point perspective, the correct viewer location lies on a horizontal semicircle whose diameter has the two vanishing points as endpoints.*

We still haven't completely solved the problem of the viewer location, and that's because we need further information. In fact, for any viewer location on the above semicircle, except for the endpoints v_1 and v_2, there is a rectangular box (a building) whose perspective image matches that in the picture (an argument for this is indicated in the exercises). However, if we know something about the actual proportions of the building, then we can pin down the correct viewer location exactly (see Exercise#). Of course, a reasonable guess is that the center of view lies at the center of the picture; this would be the case for an uncropped photograph. In this case, the correct viewpoint is the unique point E on the semicircle directly opposite the center of the picture; that is, the line through E and the center of the picture is orthogonal to the picture plane (Figure 33).

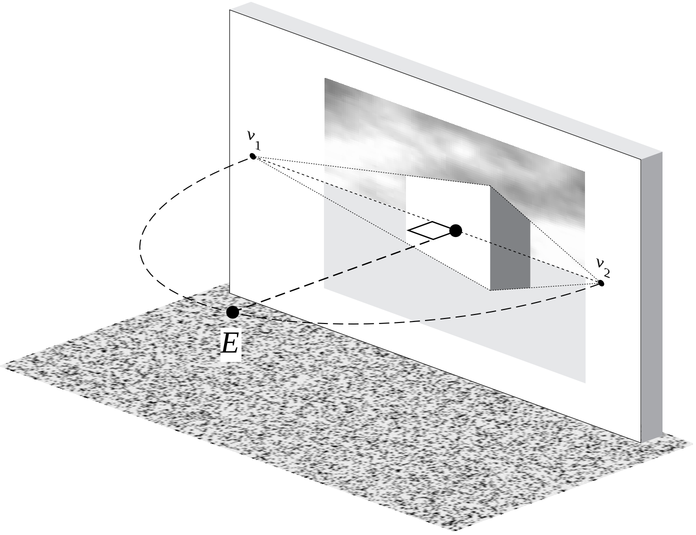


Figure 33. E is easy to locate if the center of view coincides with the center of the picture.

This gives us a method for viewing such a picture correctly. First, estimate where the two vanishing points are, and imagine the semicircle connecting them. Then walk around the semicircle until your are directly opposite the center of the picture; you should be at just about the right spot for appreciating the perspective effect intended by the artist!

Notice also that as the two vanishing points v_1 and v_2 get farther apart, the circle in Figure 33 gets larger, forcing the viewer farther away from the picture. This is important, because, as we have seen again and again, a poorly planned perspective drawing can result in a viewpoint which is uncomfortably close to the picture. The moral for two-point perspective drawings is: Keep those vanishing points far apart!

4. Measuring Points

In Section 2 we saw how the concepts of vanishing points and vanishing lines could be used in making perspective drawings. Recall, however, that we said we wouldn't worry about things like viewing distance or the size and shape of objects while constructing the drawings; that's because it made it easier to illustrate how vanishing points and vanishing planes can be applied to drawing. As a consequence, we just guessed the height of the roof peak in Figure 20 and the distance between the first two fenceposts in Figure 21. In fact, most art students do the same thing when they make their first perspective drawings; some of the decisions are done by guesswork and intuition, and others are determined by mathematically-based techniques like the use of vanishing points and vanishing planes.

In this section we'll introduce a technique which allows more accurate drawings when the size, shape, and location of objects which we wish to draw are known. **********************

5. Anamorphic Art

There are times when the correct viewer location for a painting may be surprisingly different than what we expect, and there are paintings which cannot be perceived correctly by viewing in a conventional way. One class of such paintings, called *anamorphic art,* became popular in the sixteenth century. An example of such a painting would be one whose subject could only be correctly perceived by viewing the painting in a cylindrical mirror. Another example is the painting in Figure 34, *The French Ambassadors* by Hans Holbein the younger (1497–1547). When viewed from directly in front, the painting seems perfectly normal, except for a strange, elongated object which seems to float above the floor. If we change our viewpoint to one at an oblique angle at the extreme right of the painting, the object is seen to be a human skull! The skull supposedly represents the transience of life. The point here for us is that for this painting, there is no single correct point of view.

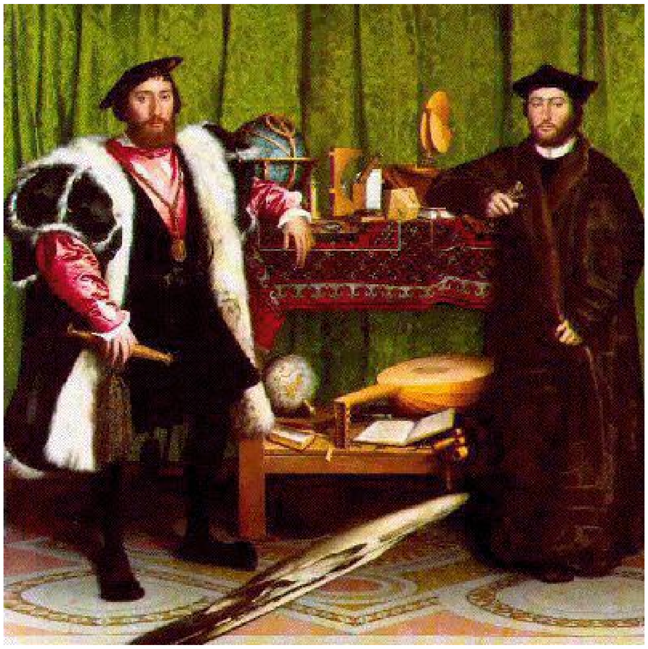


Figure 34. Hans Holbein the younger, *The French Ambassadors.*

Another example of a painting which requires an unusual viewpoint is the anamorphic portrait of Edward VI in Figure 35. In order to see the portrait correctly, we must view the painting from so far to the right of the canvas that the picture frame had to be cut away to keep from blocking our view!

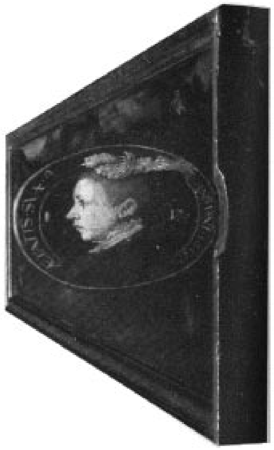


Figure 35. Unknown artist, *Anamorphic Portrait of Edward VI.* 1546.

It seems certain that such a picture, when viewed from in front, would look very strange, and indeed this the case, as can be seen in Figure 36.

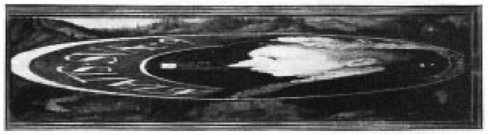


Figure 36. Frontal view of the painting in Figure 36.

How did the artist figure out how to do it? In a theoretical sense, perhaps, the trick is no big deal. The canvas is still part of the picture plane—the only new feature is that it lies completely to the left of the $x'y'$ origin. However, the human head is not usually given in xyz-coordinates, and it does not consist of planes and straight lines, so simple computation isn't enough. What is needed a combination of the artist's skill in painting portraits in the usual way, along with some sound mathematical reasoning to achieve the correct "distortion."

Let's try the trick ourselves, using a smiley face as our subject. Since our drawing will apparently be rather elongated horizontally, let's start with a canvas which is twice as wide as it is high. For later use, we'll cover it with 8 rows and 16 columns of squares (Figure 37).

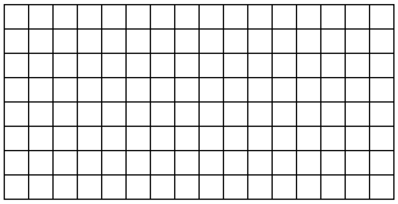


Figure 37.

Rather than try to draw the face with the correct distortion and hope it turns out right, let's imagine what the finished product should look like when seen from the right side at an oblique angle. To help with the visualization, we'll draw our canvas in perspective: the steps are shown in Figure 38.

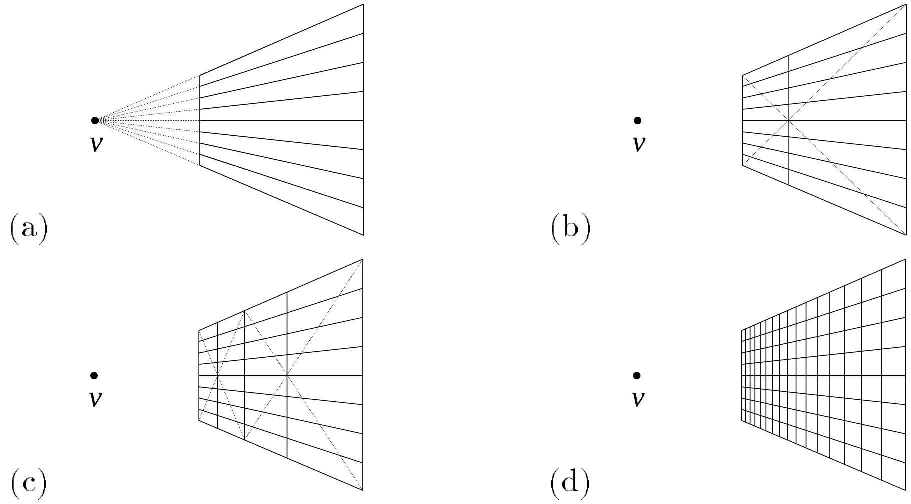


Figure 38.

In step (a) we mark a vanishing point v: we'll take this as the center of view, although this is not necessary. To draw the canvas in perspective, we make the sides vertical for convenience, and make the lines determined by the top and bottom edges converge to v. Since the sides of the canvas were located rather arbitrarily, how can we be sure that the outline of the canvas corresponds to a rectangle seen in perspective which is twice as wide as it is high? The answer is given by Theorem 3, which says that the "shape ratio" of the canvas is proportional to the viewing distance. We have chosen the shape ratio *first,* so the correct viewing distance is therefore completely determined, and from this distance the outline will appear just as it should. (You should be able to estimate this viewing distance!) To draw the horizontal lines between the squares, we make use of Theorem 2, which says that the right edge of the canvas (which is parallel to the picture plane), and all of the "features" on it, have undistorted images. The "features" we refer to are the right endpoints of the lines between the rows of squares, and thus the images of these endpoints will be equally spaced. Having located them, we then connect them with straight lines to v, which must be their vanishing point (Theorem 5).

In step (b) we locate the center vertical line of the canvas, using the same technique and reasoning that we used to locate the peak of the roof of the little house in in Figure 20 (c). In step (c) we locate the vertical lines at the 1/2 and 3/4 points of the canvas using the same technique, and continuing in this fashion we complete the process in step (d). We deliberately chose the number of columns of squares to be a power of two, since it makes this technique easy. Figure 38 (d) is now a perspective image of the gridded canvas in Figure 37.

To finish visualizing the picture, we should view the grid in Figure 38 (d) from the correct point of view and draw the smiley face on it the way we wish it to appear, which we have done in Figure 39.*

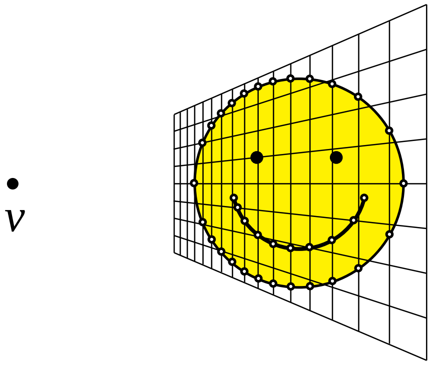


Figure 39.

Notice that we have marked many of the points where the curves of the drawing cross the grid. In order to make the actual anamorphic drawing, we now locate the corresponding version of each such point on the frontal view of the grid in Figure 40, and connect the points with smooth curves. We have mainly used intersection points on the *vertical* lines of the grid, since the proportions of features on these lines are preserved; i.e., if a point on the side of a square is, say, two-thirds of the way up the side of the square in Figure 39, then the same will be true of its counterpart in Figure 40.

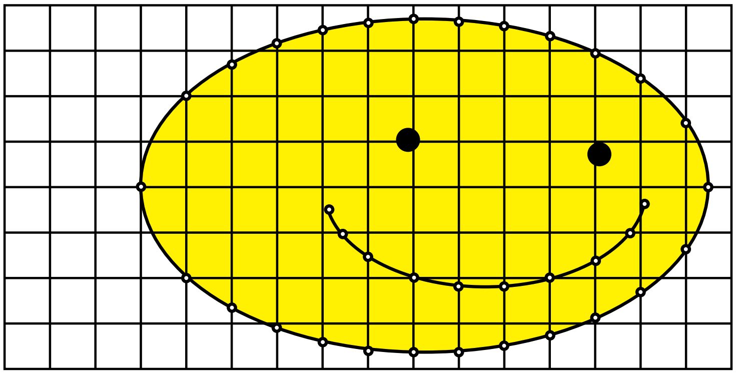


Figure 40.

If you view the enlarged anamorphic drawing in Figure 41 from the extreme right side of the page with one eye, you can see it as being undistorted. A similar technique could be used by a portrait painter to achieve effects like that in Figure 34 or Figure 35.

* There is a fine point here which we won't worry too much about. We decided that the point v would be the center of view, and hence the smiley face should not be drawn as a perfect circle in Figure 39, because then it would not *appear* to be a circle from this viewpoint. However, we have drawn it as a circle anyway, because the resulting distortion is not too great.

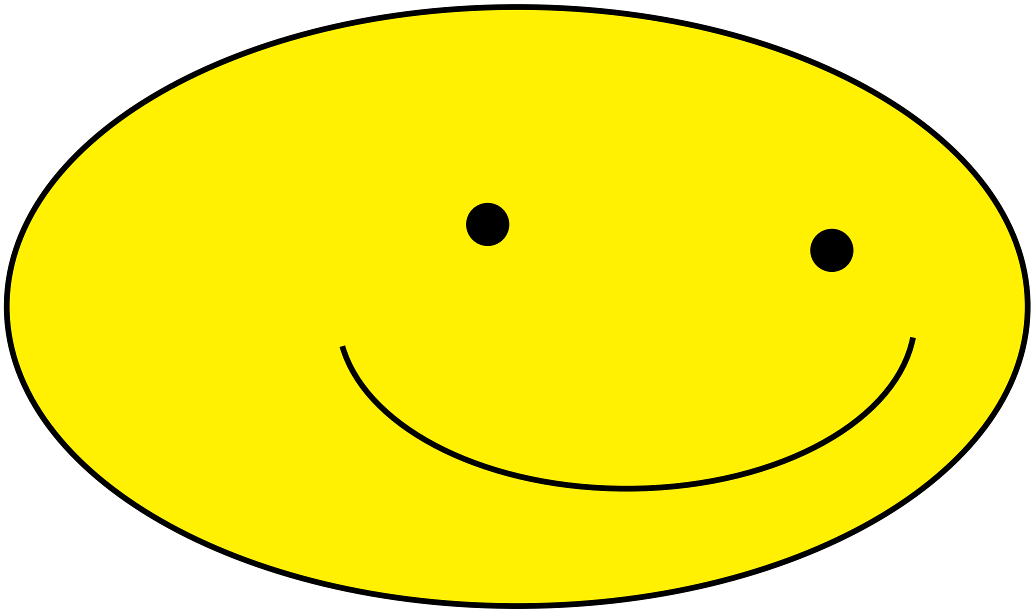


Figure 41.

Now try an anamorphic drawing yourself. Make a transparent photocopy of the perspective grid in Figure 42, and a paper photocopy of the grid in Figure 43. Place the transparent grid from Figure 42 over a simple image (such as a five-pointed star or a heart, or perhaps the face of a comic strip character), and use an erasable overhead projector pen to trace the image onto the grid. Then note where the tracing crosses the grid, mark the corresponding points on the rectangular grid from Figure 43, and connect them appropriately to make your anamorphic drawing, using color if you like. (The slight distortion mentioned in our previous footnote will again be introduced, but the result should still be good.)

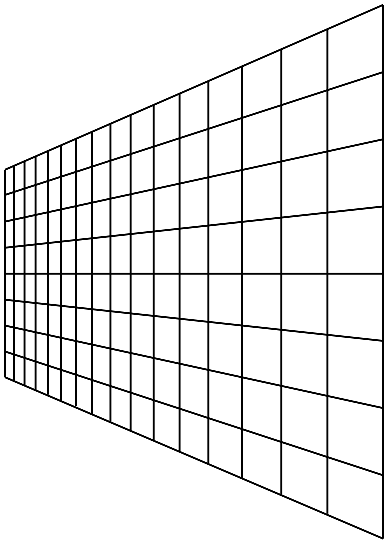


Figure 42.

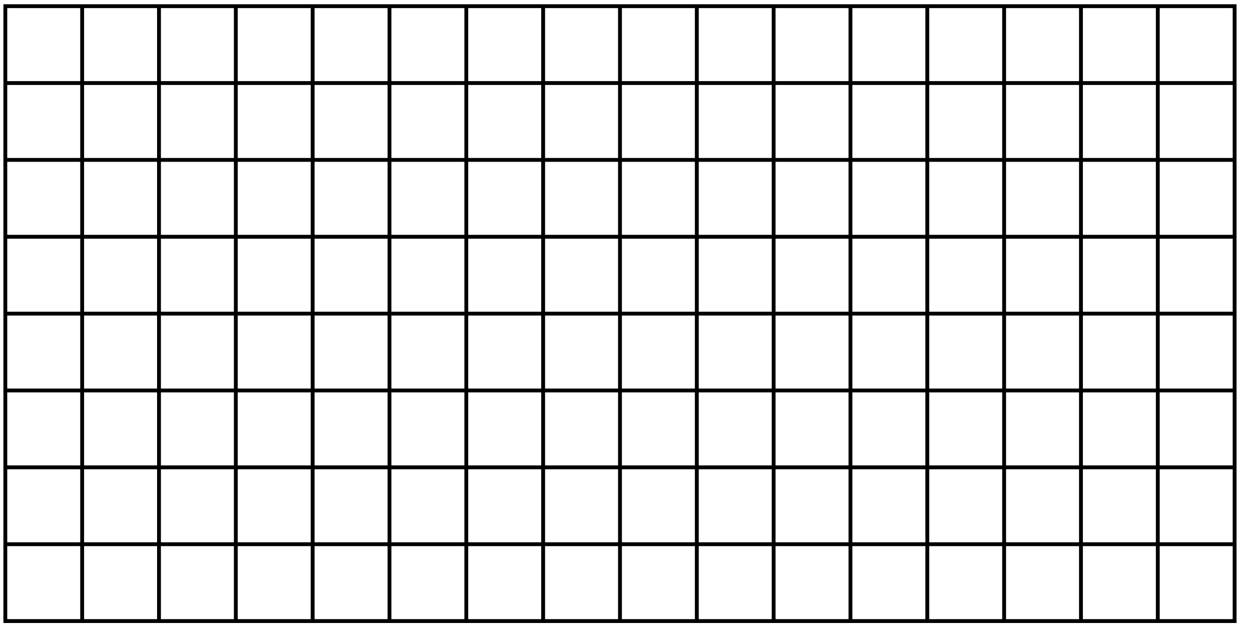


Figure 43.

Fractal Geometry

1. The Geometry of Nature

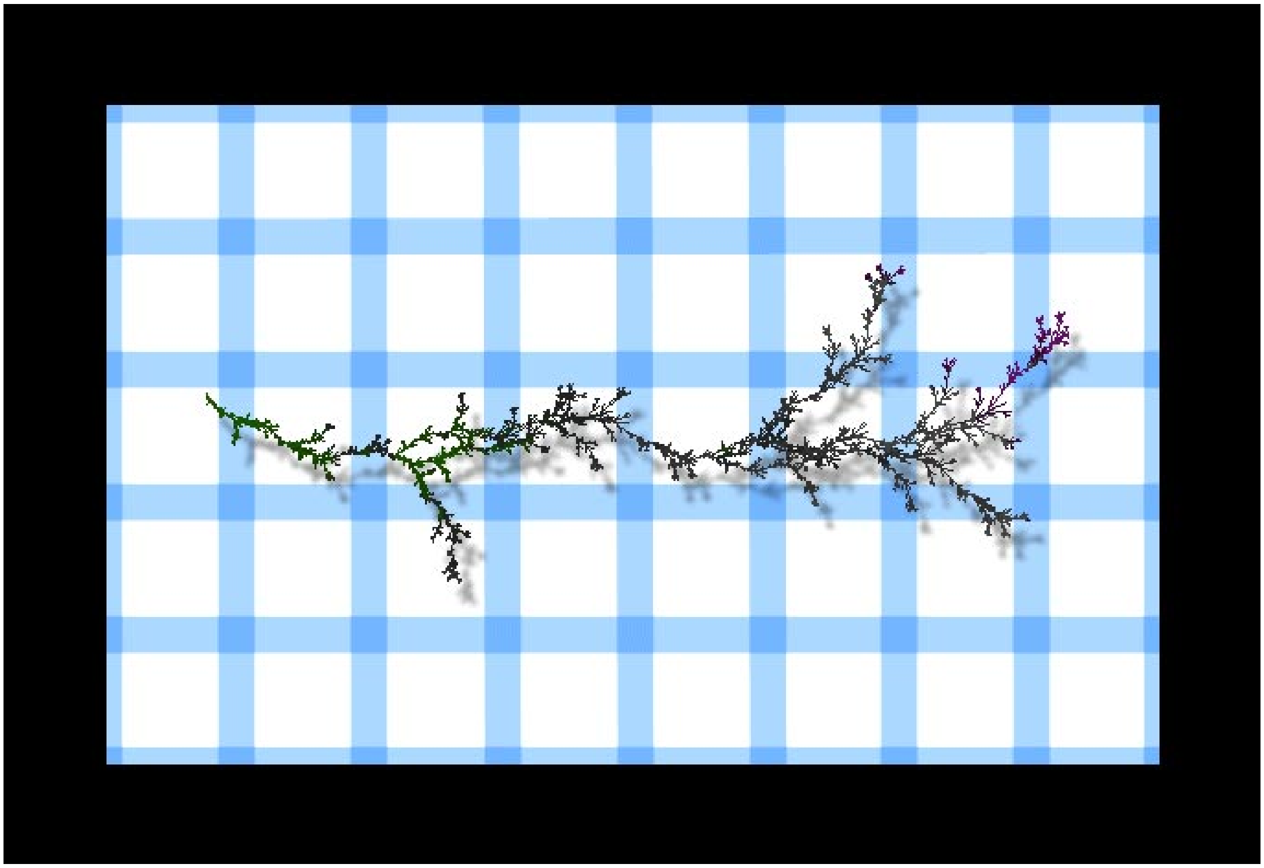


Figure 1. *Vine and Tablecloth*

The graphic entitled *Vine and Tablecloth* in Figure 1 is a visual metaphor for the contrast that many people feel exists between mathematics and the natural world. The pattern of the tablecloth reminds us of the coordinate grid of the xy-plane, with its neat rows and columns of little boxes. The vine has a completely different character. It exhibits a wild, scraggly, complex form, beautiful in a way that is both rugged and delicate. In no way does it fit into the rigid order of the coordinate grid, and it seems to defy us to describe it with mathematics. In fact, many forms in nature have this character, and this presents a problem for us. After all, we are trying to link art with mathematics, and the complex beauty of nature is one of the most important subjects of the visual artist. For example, the Chinese painting in Figure 2 pays homage to precisely this kind of wild, complex beauty, and as we will see, oriental artists have long admired those forms in nature which look least like what we traditionally think of as geometry. The subject of the ancient Chinese philosophy of Taoism is the *tao*, which can be interpreted as the “way” or the “course” of nature, and the following quote by the popular philosopher Alan Watts on the tao neatly summarizes our apparent difficulty in finding a mathematical description of the geometry of nature:

> *The tao is a certain kind of order, and this kind of order is not quite what we call order when we arrange everything geometrically in boxes, or in rows. That is a very crude kind of order, but when you look at a plant it is perfectly obvious that the plant has order. We recognize at once that is not a mess, but it is not symmetrical and it is not geometrical looking. The plant looks like a Chinese drawing, because they appreciated this kind of non-symmetrical order so much that it became an integral aspect of their painting. In the Chinese language this is called* li*, and the character for* li *means the markings in jade. It also means the grain in wood and the fiber in muscle. We could say, too, that clouds have* li*, marble has* li*, the human body has* li*. We all recognize it, and the artist copies it whether he is a landscape painter, a portrait painter, an abstract painter, or a non-objective painter. They all are trying to express the essence of* li*. The interesting thing is, that although we all know what it is, there is no way of defining it*[1].

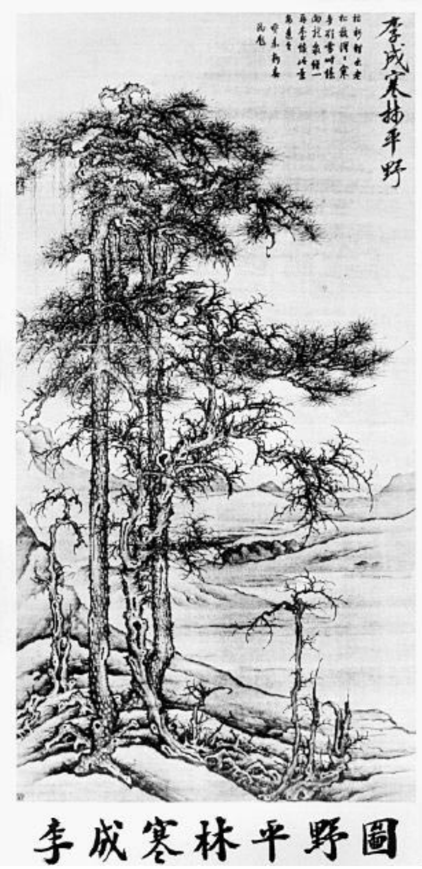


Figure 2.

At the time he wrote these words, Watts was indisputably correct. However, Alan Watts died in 1973, and that is significant, for only two years later there appeared a book entitled *Les objects fractals: forme, hazard et dimension*[2], which later evolved into an English version entitled *The Fractal Geometry of Nature*[3]. The author of these books, a mathematician named Benoit Mandelbrot, had discovered a new kind of geometry, called *fractal geometry*, which would radically change the way mathematicians and scientists—as well as many artists—viewed the natural world. Mandelbrot coined the word "fractal" from the Latin adjective *fractus*, meaning "fragmented" or "irregular," because of the forms fractal geometry describes. In his introduction to *The Fractal Geometry of Nature*, Mandelbrot echoes the sentiments of Watts concerning the discrepancies between traditional geometry and nature, but goes on to announce that the scope of "geometry" has now been widened dramatically:

> *Why is geometry often described as "cold" and "dry?" One reason lies in its inability to describe the shape of a cloud, a mountain, a coastline, or a tree. Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line.*
>
> *More generally, I claim that many patterns of Nature are so irregular and fragmented, that, compared with* Euclid*—a term used in this work to denote all of*

[1] Footnote 1.

[2] Footnote 2.

[3] Footnote 3.

standard geometry—Nature exhibits not simply a higher degree but an altogether different level of complexity. The number of distinct scales of length of natural patterns is for all practical purposes infinite.

The existence of these patterns challenges us to study those forms that Euclid leaves aside as being "formless," to investigate the morphology of the "amorphous." Mathematicians have disdained this challenge, however, and have increasingly chosen to flee from nature by devising theories unrelated to anything we can see or feel.

Responding to this challenge, I conceived and developed a new geometry of nature and implemented its use in a number of diverse fields. It describes many of the irregular and fragmented patterns around us, and leads to full-fledged theories, by identifying a family of shapes I call fractals.

Figure 1 is our first piece of evidence that Mandelbrot's bold claim is valid, because the vine in the figure is a fractal: it was drawn with mathematics! Further evidence that fractal geometry is appropriate for many natural forms can be seen by looking ahead at the figures in the text, including figures that suggest that artists have intuitively used fractal principles to depict nature for centuries. The best evidence, however, will be the beautiful, never-seen-before fractals that *you* are going to make yourself, once you understand some of the basic principles of fractal geometry. One of the best places to begin is with a drawing technique called an *iterated function system.* The artistic possibilities (as well as many other aspects) of iterated function systems were first investigated by mathematician Michael Barnsley.

2. Iterated function systems

Let's play a game that makes an interesting picture, just by plotting points in the plane. Let $f, g : \mathbb{R}^2 \to \mathbb{R}^2$ be affine transformations defined by

$$f(\mathbf{x}) = A_1\mathbf{x} + \mathbf{b}_1, \qquad \text{and} \qquad g(\mathbf{x}) = A_2\mathbf{x} + \mathbf{b}_2$$

where (1)

$$A_1 = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 2/3 & 0 \\ 0 & 2/3 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 1/3 \\ 0 \end{bmatrix}.$$

We'll also need a coin for the game, and a starting point $\mathbf{x}_0$ in the plane. It doesn't matter what starting point we use; we'll pick $\mathbf{x}_0 = [\,-1/2 \quad 0\,]^T$. Now flip the coin. If it comes up heads, we pick f, and if it comes up tails, we pick g. After each coin flip, we will plot a point in the plane. To see how the points are determined, look at this example of how a typical game might go (you should check a few of the steps to verify that the computations have been done correctly):

	COIN FLIP RESULT	CHOICE OF FUNCTION	POINT TO PLOT
Step 0	–	–	$\mathbf{x}_0 = \begin{bmatrix} -1/2 \\ 0 \end{bmatrix}$
Step 1	Heads	f	$\mathbf{x}_1 = f(\mathbf{x}_0) = \begin{bmatrix} -1/4 \\ -1/4 \end{bmatrix}$
Step 2	Tails	g	$\mathbf{x}_2 = g(\mathbf{x}_1) = \begin{bmatrix} 1/6 \\ -1/6 \end{bmatrix}$
Step 3	Tails	g	$\mathbf{x}_3 = g(\mathbf{x}_2) = \begin{bmatrix} 4/9 \\ -1/9 \end{bmatrix}$
Step 4	Heads	f	$\mathbf{x}_4 = f(\mathbf{x}_3) = \begin{bmatrix} 5/18 \\ 1/6 \end{bmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Thus we wind up plotting a sequence $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots\}$ of points in the plane. If we already have a point $\mathbf{x}_n$, we get the next point $\mathbf{x}_{n+1}$ by using a coin flip to randomly choose f or g, and then apply that function to $\mathbf{x}_n$. For instance, if we choose g, then $\mathbf{x}_{n+1} = g(\mathbf{x}_n)$. The points $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ computed above have been graphed in Figure 3. The letters f and g have been used to show which function was used to "move" the point $\mathbf{x}_n$ to the next point $\mathbf{x}_{n+1}$.

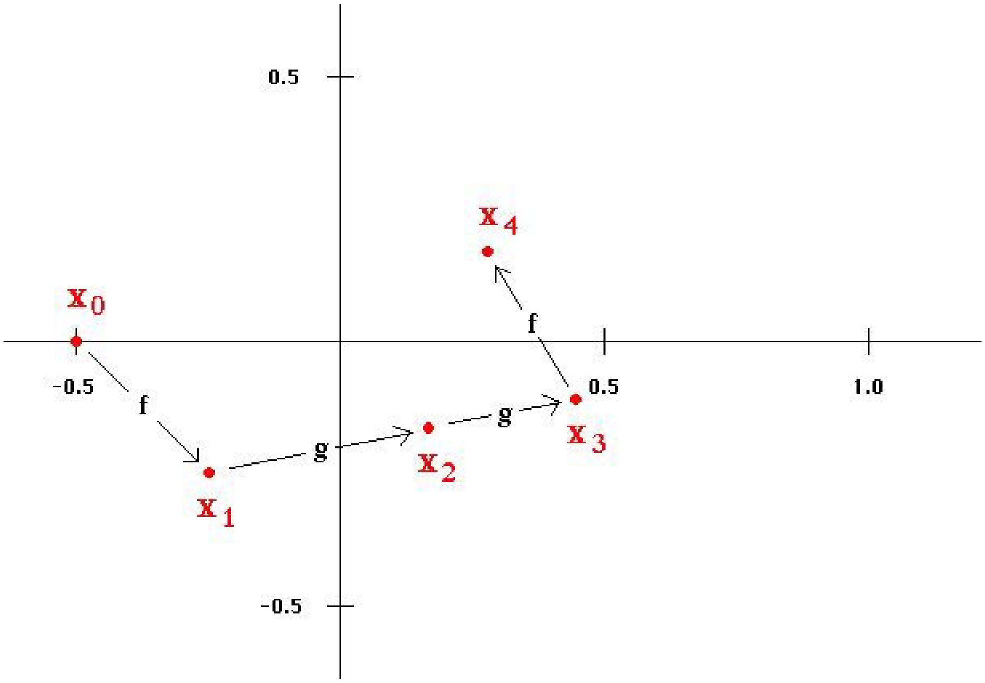


Figure 3.

So far it doesn't look like much—just a random-looking collection of dots. However, look at Figure 4 to see what we get if we use a computer to continue the process!

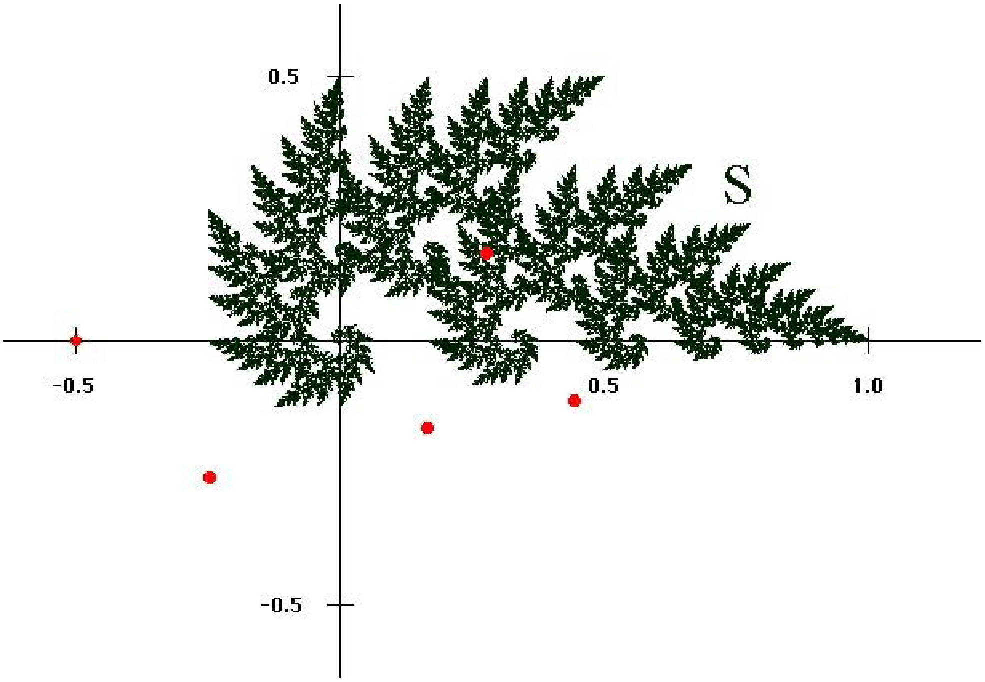


Figure 4.

In Figure 4 we have reproduced the first five points $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ in red, and the next 80,000 or so points in green. The green points make a fantastic image that seems to be made of infinitely repeated pine trees! The set S indicated by the green points is, of course, a fractal, and the game we played to draw it is called an *iterated function system.* To iterate means to do something repeatedly—in this case, to repeatedly get new points by applying f or g to the previous point. When a fractal is made with an iterated function system, or IFS, the fractal is called the *attractor* of the IFS. You can see why by noticing that, although the points $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ do not lie on S, they seem to be getting closer to it (refer to the labels in Figure 3), and in fact, the point $\mathbf{x}_4$ seems to lie on—or very nearly on—the attracting set S. After that, the green points $\mathbf{x}_5, \mathbf{x}_6, \ldots$ fill out the attractor so we can see it. The fractal in Figure 1 can be created in a similar way, and we'll show how to draw it later.

In Figure 4 we have created an infinitely complex "organic" shape with just two very simple affine transformations; such a seemingly miraculous feat needs some explaining! However, before we begin trying to understand this process, the are some points we should make absolutely clear:

- The range of shapes attainable with fractal geometry is endless: you can make trees, plants, clouds, mountains, galaxies, flowers, snowflakes, lightning, rivers, coastlines, and many other natural forms, as well as endless beautiful and strange "abstract" shapes and textures. The artistic possibilities are awsome!

- Despite the infinite complexity of fractal forms, the basic principles of fractal geometry are *easy* to understand and implement creatively.

- Understanding fractal geometry does *not* make the magic go away. Whether you study fractal geometry for weeks or years, the fascination of seeing fractal shapes appear never diminishes, and there are always surprises.

- Understanding fractal geometry will help you understand and draw Nature more effectively. However, this understanding of natural forms does *not* make Nature seem technical, boring, or less magical. You cannot "kill" the beauty and mystery of Nature by achieving greater understanding of it! In fact, it is well-known by scientists that the *opposite* effect occurs: the realization of order amid complexity in Nature makes the contemplation and experience of it so touching and awe-inspiring that the effect is addictive! This is a Great Well-Kept Secret of Science and Mathematics: that the process of understanding *deepens* and *multiplies* the mysteries. Fractal geometry allows the ordinary person to take part in this process of wonder, in an authentic and satisfying way.

So now that we have that straight, let's start investigating the process by which the image in Figure 4 was drawn. First, let's look more closely at the functions f and g defined in Equations (1). The matrix A_1 is actually a scalar multiple of a rotation matrix. To see this, observe that

$$A_1 = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix} = \sqrt{2}/2 \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} = \sqrt{2}/2 \begin{bmatrix} \cos 45^\circ & -\sin 45^\circ \\ \sin 45^\circ & \cos 45^\circ \end{bmatrix},$$

so A_1 rotates vectors 45° counterclockwise, and shrinks them by a factor of $\sqrt{2}/2$, or approximately 0.7. Since $\mathbf{b}_1 = \mathbf{0}$, we have $f(\mathbf{x}) = A_1\mathbf{x}$ for each $\mathbf{x} \in \mathbb{R}^2$, so f also rotates vectors 45° counterclockwise, and shrinks them by a factor of ≈ 0.7. It should be clear from the definition of g in (1) that g shrinks vectors by a factor of $2/3$ and translates them $1/3$ unit to the right. Let's summarize these results by saying what f and g do to *subsets* of $\mathbb{R}^2$:

(i) f rotates sets 45° counterclockwise about the origin, and shrinks them toward the origin by a factor of $\sqrt{2}/2 \approx 0.7$;

(ii) g shrinks sets toward the origin by a factor of $2/3$, and then translates the resulting sets $1/3$ unit to the right.

It would be nice to see an example of f and g doing these things to a subset of $\mathbb{R}^2$, and for our purposes, the best example is the set S itself! This is illustrated in Figure 5, where S is the entire red-and-blue set, $f(S)$ is the red set, and $g(S)$ is the blue set. That is, if we apply f to every point $\mathbf{x}$ of S, and color the resulting points $f(\mathbf{x})$ red, we get the set $f(S)$ in Figure 5, and similarly for $g(S)$, which is blue. In terms of the descriptions (i) and (ii), the set $f(S)$ is what we get when we rotate the set S through an angle of 45° counterclockwise about the origin, and shrink it toward the origin by a factor of $\sqrt{2}/2 \approx 0.7$. The set $g(S)$ is what we get when we shrink the set S toward the origin by a factor of $2/3$, and then translate the resulting set $1/3$ unit to the right. Apparently, a very unusual thing happens: the set S is precisely the union of $f(S)$ and $g(S)$. That is, apparently,

$$S = f(S) \cup g(S).$$

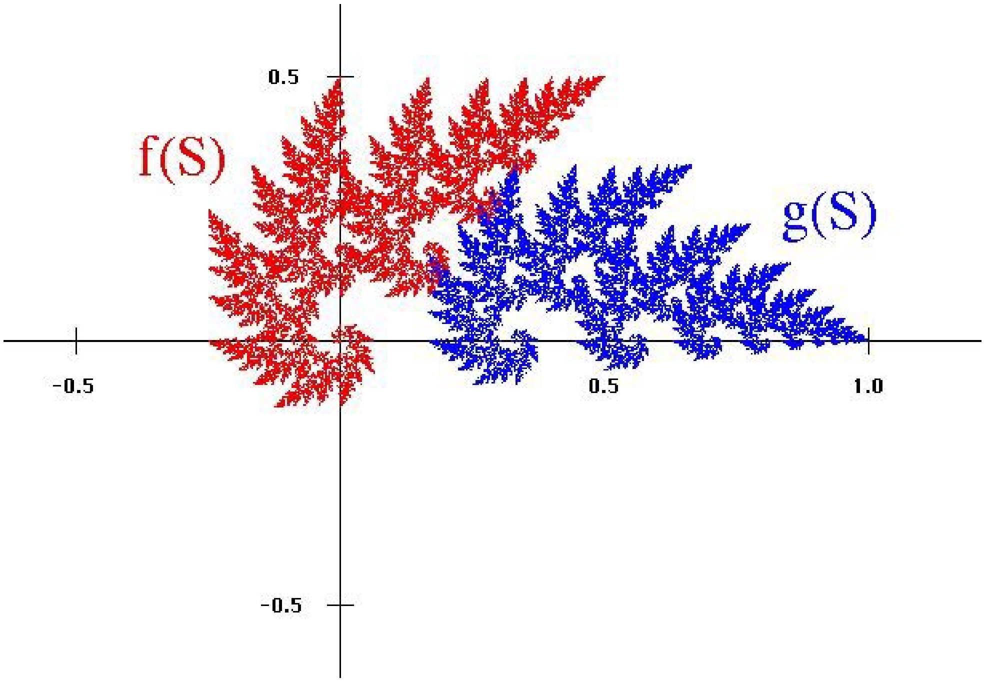


Figure 5. The set S has the property that $S = f(S) \cup g(S)$.

We'll show that this is indeed the case. In fact, this situation is related to one of the most important properties of fractals, sometimes referred to as "self-similarity." We'll say more about self-similarity later, but basically it means that "the parts look like the whole."

Before going into that, let's take care of some other business. In order to be good at drawing pictures with iterated function systems, let's precisely

I. Define what we mean by an iterated function system (IFS);

II. Define the attractor S of an IFS;

III. Define what it means for the attractor to "attract" points, and show that it does so;

IV. Show that in the example above, $S = f(S) \cup g(S)$, and generalize the result to other IFS's.

Notice that the affine maps f and g were contraction mappings on $\mathbb{R}^2$. Specifically, from what we have said it follows that $||A_1|| = \sqrt{2}/2$ and $||A_2|| = 2/3$. These are the kinds of mappings which we will use in our iterated function systems. Since we used a coin to choose f or g, each map had a 1/2 probability of being chosen. More generally, we may use more than just two mappings, and we will allow our maps to have any nonzero probability of being chosen, the only restriction being that the probabilities of all the maps must add up to 1, so that it's absolutely certain that *some* map will be chosen on each turn. Although more general types of iterated function systems exist, the IFS's we will use here are capable of generating a fantastically rich variety of fractals. Our definition is as follows.

I. An *iterated function system* (IFS) is a collection of affine contraction mappings

$$f_1(\mathbf{x}) = A_1\mathbf{x} + \mathbf{b}_1,\ f_2(\mathbf{x}) = A_2\mathbf{x} + \mathbf{b}_2,\ \ldots,\ f_n(\mathbf{x}) = A_n\mathbf{x} + \mathbf{b}_n$$

on $\mathbb{R}^2$, with corresponding nonzero probabilities

$$p_1, p_2, \ldots, p_n \qquad \text{such that} \qquad p_1 + p_2 + \cdots + p_n = 1.$$

We described the attractor S in Figure 4 as being "infinitely complex," and we'll see that this is true in a sense. It may seem impossible to precisely define such a set, but in fact, it's not hard to do. Before giving the definition, however, let's discuss our example with f and g a little further, so that we can see where the idea for the definition comes from. We claimed that the sequence of points $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ was being "attracted" by the set S. If we look at this sequence in terms of the matrices A_1, A_2 and the vectors $\mathbf{b}_1$, $\mathbf{b}_2$, we see that

$$\begin{aligned}
\mathbf{x}_1 &= f(\mathbf{x}_0) = A_1\mathbf{x}_0 + \mathbf{b}_1\\
&= \mathbf{b}_1 + A_1\mathbf{x}_0\\
\mathbf{x}_2 &= g(\mathbf{x}_1) = A_2(\mathbf{b}_1 + A_1\mathbf{x}_0) + \mathbf{b}_2\\
&= \mathbf{b}_2 + A_2\mathbf{b}_1 + A_2A_1\mathbf{x}_0\\
\mathbf{x}_3 &= g(\mathbf{x}_2) = A_2(\mathbf{b}_2 + A_2\mathbf{b}_1 + A_2A_1\mathbf{x}_0) + \mathbf{b}_2\\
&= \mathbf{b}_2 + A_2\mathbf{b}_2 + A_2A_2\mathbf{b}_1 + A_2A_2A_1\mathbf{x}_0\\
\mathbf{x}_4 &= f(\mathbf{x}_3) = A_1(\mathbf{b}_2 + A_2\mathbf{b}_2 + A_2A_2\mathbf{b}_1 + A_2A_2A_1\mathbf{x}_0) + \mathbf{b}_1\\
&= \mathbf{b}_1 + A_1\mathbf{b}_2 + A_1A_2\mathbf{b}_2 + A_1A_2A_2\mathbf{b}_1 + A_1A_2A_2A_1\mathbf{x}_0\\
&\ \ \vdots
\end{aligned}$$

We said earlier that the point $\mathbf{x}_4$ seemed to be very close to the set S, so let's look at the expression for $\mathbf{x}_4$ more closely to see if we can guess what the formula for a "typical point" of S should look like. Equation (2) gives the expression for $\mathbf{x}_4$ in color-coded form:

$$\mathbf{x}_4 = \mathbf{b}_1 + A_1\mathbf{b}_2 + A_1A_2\mathbf{b}_2 + A_1A_2A_2\mathbf{b}_1 + A_1A_2A_2A_1\mathbf{x}_0\,. \tag{2}$$

The magenta symbols came from the first application of f to $\mathbf{x}_0$; the blue symbols came from the subsequent application of g to the point $\mathbf{x}_1 = f(\mathbf{x}_0)$; the green symbols came from the application of g to the point $\mathbf{x}_2 = g(\mathbf{x}_1)$; and the red symbols came from the final application of f to the point $\mathbf{x}_3 = g(\mathbf{x}_2)$. If we continue this process, we'll run out of colors, so let's replace colors with subscripts. If the red symbols are given new subscripts k_1, the green symbols given new subscripts k_2, the blue symbols given new subscripts k_3, and the magenta symbols given subscripts k_4, then Equation (2) becomes

$$\mathbf{x}_4 = \mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + A_{k_1}A_{k_2}A_{k_3}\mathbf{b}_{k_4} + A_{k_1}A_{k_2}A_{k_3}A_{k_4}\mathbf{x}_0. \tag{3}$$

Notice that the subscript k_1 is associated with the *last* function used, not the first one, so if we applied another function to $\mathbf{x}_4$, we'd have to re-number all the subscripts. However, if we keep things in this more general form, it's easy to write down the form of a generic point $\mathbf{x}_N$:

$$\begin{aligned}\mathbf{x}_N =& \mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + \cdots + A_{k_1}A_{k_2}\cdots A_{k_{N-1}}\mathbf{b}_{k_N}\\ &+ A_{k_1}A_{k_2}\cdots A_{k_N}\mathbf{x}_0.\end{aligned}$$

If S really attracts points as we have said, then for large N, the point $\mathbf{x}_N$ should be close to some point of S. This suggests that a generic point $\mathbf{y}$ of S should be defined by a formula that looks something like the formula for $\mathbf{x}_N$. However, we implied earlier that the "attraction" takes place regardless of the starting point $\mathbf{x}_0$, so we shouldn't use $\mathbf{x}_0$ in the formula for $\mathbf{y}$. Moreover, there is no logical number of terms to stop with, so let's not stop; let's try writing $\mathbf{y}$ as an infinite series of vectors $\mathbf{b}_{k_j}$, all except the first one multiplied by matrices A_{k_l}:

$$\mathbf{y} = \mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + A_{k_1}A_{k_2}A_{k_3}\mathbf{b}_{k_4} + \cdots, \tag{4}$$

where each subscript k_j is either a 1 or a 2; that is, each vector $\mathbf{b}_{k_j}$ is either $\mathbf{b}_1$ or $\mathbf{b}_2$, and each matrix A_{k_j} is either A_1 or A_2. It's important to realize that if, say, A_{k_5} represents the matrix A_2 (that is, if $k_5 = 2$), then A_{k_5} represents A_2 everywhere it appears, and furthermore, $\mathbf{b}_{k_5}$ must then represent $\mathbf{b}_2$. We have achieved what we were after—a reasonable guess at a formula that describes a "typical" element of S. Of course, the series in (4) will not be meaningful unless it converges, but we can see that it does, by looking at the corresponding series of norms:

$$\|\mathbf{b}_{k_1}\| + \|A_{k_1}\mathbf{b}_{k_2}\| + \|A_{k_1}A_{k_2}\mathbf{b}_{k_3}\| + \|A_{k_1}A_{k_2}A_{k_3}\mathbf{b}_{k_4}\| + \cdots. \tag{5}$$

In this series, the Nth term, for $N \geq 2$, is $\|A_{k_1}A_{k_2}\cdots A_{k_{N-1}}\mathbf{b}_{k_N}\|$. If we set $\alpha = \max\{\|A_1\|, \|A_2\|\}$, $\beta = \max\{\|\mathbf{b}_1\|, \|\mathbf{b}_2\|\}$, we see that the Nth term satisfies

$$\|A_{k_1}A_{k_2}\cdots A_{k_{N-1}}\mathbf{b}_{k_N}\| \leq \alpha^{N-1}\beta,$$

so the series in (5) is dominated by the geometric series

$$\beta + \alpha\beta + \alpha^2\beta + \cdots,$$

which converges, since $0 < \alpha = \sqrt{2}/2 < 1$. It follows that the series in (4) converges also, and we are now ready to give our definition of the attractor of an IFS.

II. Let an iterated function system be given, with affine contraction mappings

$$f_1(\mathbf{x}) = A_1\mathbf{x} + \mathbf{b}_1,\ f_2(\mathbf{x}) = A_2\mathbf{x} + \mathbf{b}_2,\ \ldots,\ f_n(\mathbf{x}) = A_n\mathbf{x} + \mathbf{b}_n$$

on $\mathbb{R}^2$. The *attractor* S of the system is the set of all points of the form $\mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + \cdots$, where each matrix A_{k_j}, and each vector $\mathbf{b}_{k_j}$, is one of those listed above. That is,

$$S = \{\mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + \cdots \mid 1 \le k_j \le n \text{ for all } j\}.$$

Now let's show that the attractor S really does "attract" points. Suppose we have an iterated function system like that given in Item I, and suppose we have a starting point $\mathbf{x}_0 \in \mathbb{R}^2$. We begin creating a sequence of points $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ by applying the functions f_k, choosing them at random according to their probabilities p_k. At the Nth step of the process, we have a point

$$\begin{aligned}\mathbf{x}_N = &\mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + \cdots + A_{k_1}A_{k_2}\cdots A_{k_{N-1}}\mathbf{b}_{k_N}\\ &+ A_{k_1}A_{k_2}\cdots A_{k_N}\mathbf{x}_0.\end{aligned}$$

We would like to show that such points eventually stay very close to points of the attractor S. Specifically, if someone picks a small distance $\varepsilon > 0$ and says, "After some step M, I want each point $\mathbf{x}_N$, for $N > M$, to be less than ε units away from some point on the attractor S." Can we do it? Yes, and here's how. Following our pervious argument, we define

$$\alpha = \max\{\|A_1\|, \|A_2\|, \ldots, \|A_n\|\} \quad \text{and} \quad \beta = \max\{\|\mathbf{b}_1\|, \|\mathbf{b}_2\|, \ldots, \|\mathbf{b}_n\|\},$$

and notice that since $0 < \alpha < 1$, the geometric series

$$\beta + \alpha\beta + \alpha^2\beta + \cdots$$

converges. Thus we can choose a positive integer M large enough so that

$$\alpha^N\beta + \alpha^{N+1}\beta + \cdots < \varepsilon/2$$

for every $N > M$. We can also insure that M is large enough so that

$$\alpha^N \|\mathbf{x}_0\| < \varepsilon/2$$

for every $N > M$.

Now suppose that $N > M$, and consider the point

$$\begin{aligned}\mathbf{x}_N = \mathbf{b}_{k_1} &+ A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + \cdots + A_{k_1}A_{k_2}\cdots A_{k_{N-1}}\mathbf{b}_{k_N} \\ &+ A_{k_1}A_{k_2}\cdots A_{k_N}\mathbf{x}_0\end{aligned}$$

and the point $\mathbf{y}$ in S defined by

$$\begin{aligned}\mathbf{y} = \mathbf{b}_{k_1} &+ A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + \cdots + A_{k_1}A_{k_2}\cdots A_{k_{N-1}}\mathbf{b}_{k_N} \\ &+ A_{k_1}A_{k_2}\cdots A_{k_N}\mathbf{b}_{k_{N+1}} + A_{k_1}A_{k_2}\cdots A_{k_{N+1}}\mathbf{b}_{k_{N+2}}\cdots\end{aligned}$$

in S. Notice that the first N terms in the expansions of $\mathbf{x}_N$ and $\mathbf{y}$ are the same. Thus they cancel when we subtract $\mathbf{y} - \mathbf{x}_N$, so the distance between $\mathbf{y}$ and $\mathbf{x}_N$ is

$$\begin{aligned}\|\mathbf{y} - \mathbf{x}_N\| = \| &- A_{k_1}A_{k_2}\cdots A_{k_N}\mathbf{x}_0 \\ &+ A_{k_1}A_{k_2}\cdots A_{k_N}\mathbf{b}_{k_{N+1}} + A_{k_1}A_{k_2}\cdots A_{k_{N+1}}\mathbf{b}_{k_{N+2}} + \cdots \| \\ \leq \| &- A_{k_1}A_{k_2}\cdots A_{k_N}\mathbf{x}_0\| \\ &+ \|A_{k_1}A_{k_2}\cdots A_{k_N}\mathbf{b}_{k_{N+1}}\| + \|A_{k_1}A_{k_2}\cdots A_{k_{N+1}}\mathbf{b}_{k_{N+2}}\| + \cdots \\ < \varepsilon&/2 + (\alpha^N\beta + \alpha^{N+1}\beta + \cdots) \\ < \varepsilon&/2 + \varepsilon/2 = \varepsilon,\end{aligned}$$

which is what we wanted to show. Therefore, we can say that

> III. For any starting point $\mathbf{x}_0$ and any $\varepsilon > 0$, there exists a positive integer M, such that for all $N > M$, each of the points $\mathbf{x}_N$ is less than ε units from some point of S. This is the sense in which the attractor S "attracts" the points $\mathbf{x}_N$.

Thus, if $N > M$, then $\mathbf{x}_N$ is within ε units of some point $\mathbf{y}$ in S, and $\mathbf{x}_{N+1}$ is within some point $\mathbf{z}$ in S, and so on. However, $\mathbf{y}$ and $\mathbf{z}$ may be far apart from each other! That is, $\mathbf{x}_{N+1}$ may jump clear to the other side of the attractor from $\mathbf{x}_N$, but nevertheless, $\mathbf{x}_N$ and $\mathbf{x}_{N+1}$ will each be close to *some* point of S. This is very important from an artistic point of view, since it means that the points $\mathbf{x}_N$ will eventually be indistinguishable from points of S, the set we want to draw.

However, there is another artistic/mathematical issue we have not taken care of. What we have said so far is like saying that if we send a group of bees (the points $\mathbf{x}_N$ for large enough N) into a flower garden (the attractor S), then every bee will be close to a flower (a point of S). But will every flower be close to a bee? Maybe not: even if we have lots of bees, maybe some bees will huddle around the same flower, and not distribute themselves nicely around the garden. Analogously, maybe the "dots" $\mathbf{x}_N$ will not "fill out" the attractor S and give us a complete picture. However, each point $\mathbf{y}$ of S eventually *will* have some point $\mathbf{x}_N$ close to it—in fact, as close as we want. To see why, suppose that

$$\begin{aligned}\mathbf{y} = \mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + \cdots + A_{k_1}A_{k_2}\cdots A_{k_{N-1}}\mathbf{b}_{k_N}\\ + A_{k_1}A_{k_2}\cdots A_{k_N}\mathbf{b}_{k_{N+1}} + A_{k_1}A_{k_2}\cdots A_{k_{N+1}}\mathbf{b}_{k_{N+2}}\cdots,\end{aligned}$$

is a point of S, and suppose we want the series expansion of some point $\mathbf{x}_N$ to agree with the first three terms of $\mathbf{y}$. Since the functions f_k are being chosen at random, eventually three of them will be chosen in the order f_{k_3}, f_{k_2}, f_{k_1}. We're assuming that our iterated function game is already in progress, so we already have gone through Q steps, and we are working on a point $\mathbf{x}_Q$. Going through three more steps, we can write $N = Q + 3$, and we have

$$\begin{aligned}\mathbf{x}_N &= f_{k_1}(f_{k_2}(f_{k_3}(\mathbf{x}_Q)))\\ &= \mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + A_{k_1}A_{k_2}A_{k_3}\mathbf{x}_Q,\end{aligned}$$

where the first three terms are the same as those in the expansion of $\mathbf{y}$. A similar argument can be made to show that we will eventually get a point which agrees with $\mathbf{y}$ in the first thousand terms, or the first million terms, etc. Then, by using an argument like the one we used to verify Item III, we can show that such points will get arbitrarily close to $\mathbf{y}$.

Finally, let's take care of Item IV.

IV. The attractor S of the iterated function system with affine contraction mappings

$$f_1(\mathbf{x}) = A_1\mathbf{x} + \mathbf{b}_1,\ f_2(\mathbf{x}) = A_2\mathbf{x} + \mathbf{b}_2,\ \ldots,\ f_n(\mathbf{x}) = A_n\mathbf{x} + \mathbf{b}_n$$

satisfies

$$S = f_1(S) \cup f_2(S) \cup \cdots \cup f_n(S).$$

To show that $S = f_1(S) \cup f_2(S) \cup \cdots \cup f_n(S)$, we must show that each element of S is an element of $f_1(S) \cup f_2(S) \cup \cdots \cup f_n(S)$, and that each element of $f_1(S) \cup f_2(S) \cup \cdots \cup f_n(S)$ is an element of S. To begin, let

$$y = \mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + \cdots$$

be an element of S. By our definition of S, the point

$$\mathbf{z} = \mathbf{b}_{k_2} + A_{k_2}\mathbf{b}_{k_3} + A_{k_2}A_{k_3}\mathbf{b}_{k_4} + \cdots$$

is also an element of S, and since we can apply the matrix A_1 term-by-term to $\mathbf{z}$, we have

$$f_{k_1}(\mathbf{z}) = A_{k_1}\mathbf{z} + \mathbf{b}_{k_1} = \mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + A_{k_1}A_{k_2}A_{k_3}\mathbf{b}_{k_4} + \cdots = \mathbf{y},$$

so $\mathbf{y}$ belongs to $f_{k_1}(S)$, and therefore it belongs to $f_1(S) \cup f_2(S) \cup \cdots \cup f_n(S)$.

Going the other way, suppose $\mathbf{y}$ belongs to $f_1(S) \cup f_2(S) \cup \cdots \cup f_n(S)$. Then $\mathbf{y}$ belongs to $f_k(S)$ for some k. We can label this k as k_1, so that $\mathbf{y} \in f_{k_1}(S)$, which means that $\mathbf{y} = f_{k_1}(\mathbf{z})$ for some $\mathbf{z} \in S$. It follows that $\mathbf{z}$ can be written in the form

$$\mathbf{z} = \mathbf{b}_{k_2} + A_{k_2}\mathbf{b}_{k_3} + A_{k_2}A_{k_3}\mathbf{b}_{k_4} + \cdots,$$

so

$$\mathbf{y} = f_{k_1}(\mathbf{z}) = A_{k_1}\mathbf{z} + \mathbf{b}_{k_1} = \mathbf{b}_{k_1} + A_{k_1}\mathbf{b}_{k_2} + A_{k_1}A_{k_2}\mathbf{b}_{k_3} + A_{k_1}A_{k_2}A_{k_3}\mathbf{b}_{k_4} + \cdots,$$

which is an element of S, by the definition of S. This finishes the proof.

For our purposes, Item IV is perhaps the most important item of all. Items I–III are important, too, because they guarantee that we have a dependable way of drawing fractals; the artistic equivalent would be making sure that we have reliable paints and brushes. However, Item IV is what allows us to understand the artwork itself. To see how, recall that the interpretation of Item IV that is appropriate to the set S in Figure 5 (reproduced in Figure 6 below) is the statement

$$S = f(S) \cup g(S).$$

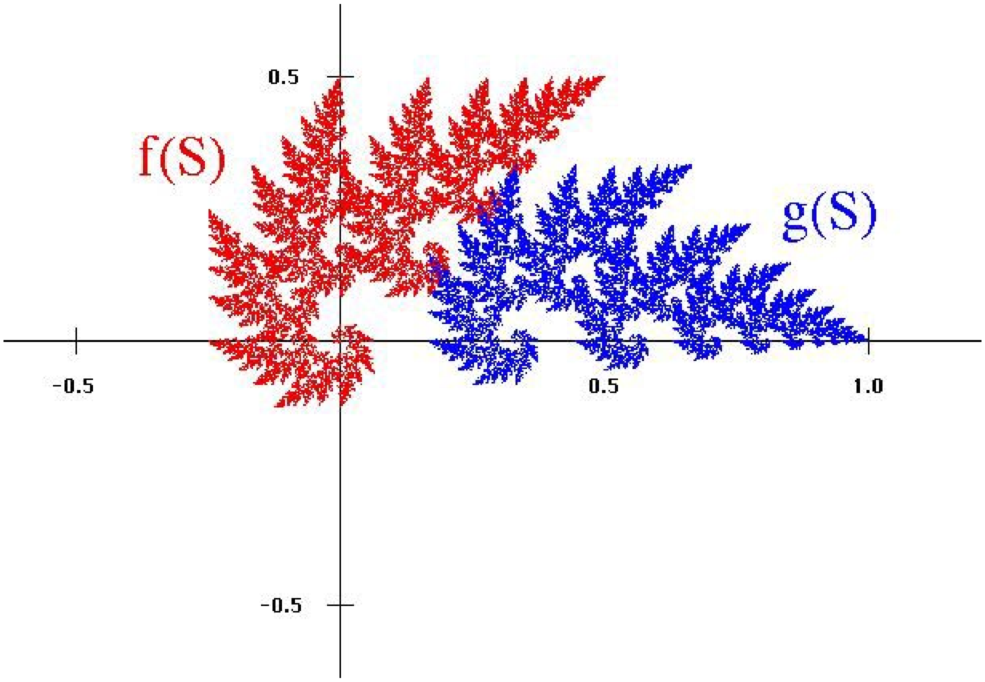


Figure 6.

Now think about the set $f(S)$. Since $S = f(S) \cup g(S)$, we can write $f(S) = f(f(S) \cup g(S))$. Thus we get $f(S)$ by applying f to every point of $f(S)$ itself, [thereby obtaining $f(f(S))$], and by applying f to every point in $g(S)$ [thereby obtaining $f(g(S))$]. It follows that $f(S)$ is the union of the two sets $f(f(S))$ and $f(g(S))$:

$$f(S) = f(f(S)) \cup f(g(S)).$$

This situation is illustrated in Figure 7, in which $f(f(S))$ is shaded in red and $f(g(S))$ is shaded in green.

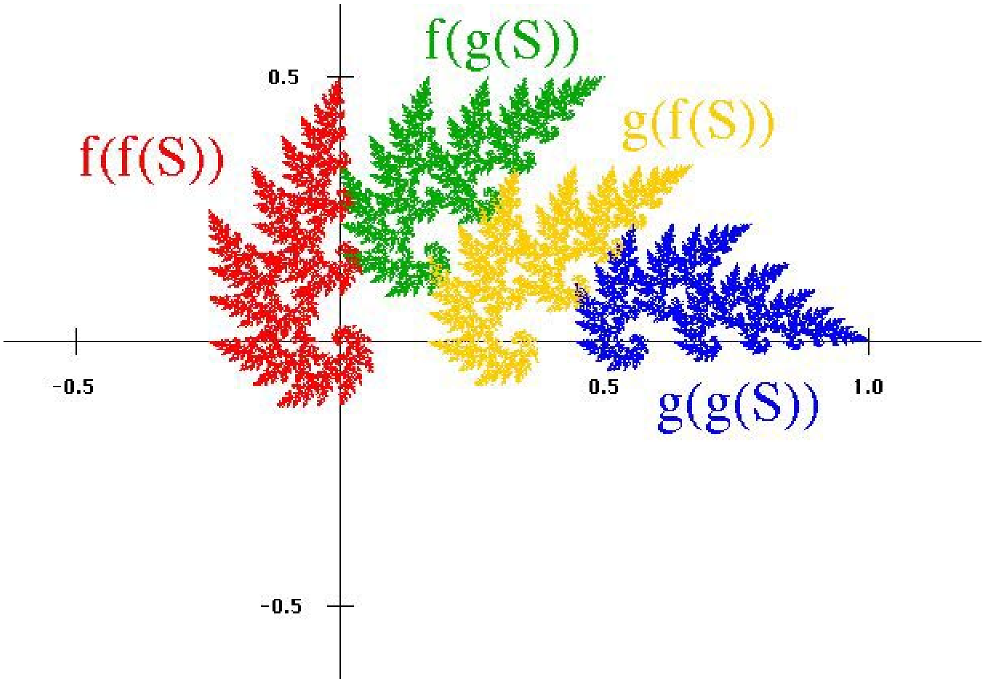


Figure 7.

In a similar manner, we can deduce that the set $g(S)$ is also a union of two sets:

$$g(S) = g(f(S)) \cup g(g(S)).$$

The sets $g(f(S))$ and $g(g(S))$ have been shaded in gold and blue, respectively. We can keep going like this, breaking the set S down into smaller and smaller parts. For instance, can you take a pencil and circle the part of S which is $f(f(f(S)))$?

All of this subdividing of S is more than just a mathematical exercise. It can help us answer questions such as

- Why does the set S look so complicated?
- Why does the set S look like it is made of trees?

The set S looks complicated because it consists of smaller and smaller copies of itself. If such a set has any interesting detail at all, then that detail will be repeated at smaller and smaller scales, without end. This is why we were justified in referring to Figure 4 as being “infinitely complex.”

The reason that the set S looks like it is made of trees is related to the same property: S consists of smaller and smaller copies of itself! If you cut a branch off a tree (remove a subset of the tree) and stick it in the ground, it looks like a tree. If you cut a smaller branch off that branch and stick it in the ground, it also looks like a tree. Of course, this process cannot go on forever, because, for example, the tree does not consist of little trees at the microscopic scale; it consists of cells, which don’t look anything like trees. However, to a good approximation, a tree consists of smaller copies of itself, so it is not too surprising that a fractal can look like a tree. In fact, you can make all kinds of fractal trees. As an example, Figure 8 features a fractal tree S which is the attractor of a five-map IFS (the origin is at the base of the tree).

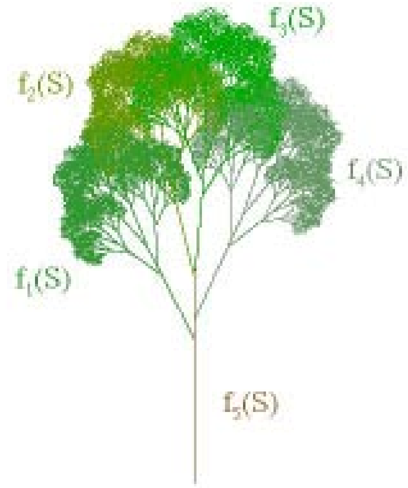


Figure 8.

Below are the affine maps which make the tree. We use the usual notation $f_k(\mathbf{x}) = A_k\mathbf{x} + \mathbf{b}_k$, and denote the corresponding probability of being chosen by p_k.

$$f_1 : \quad A_1 = \begin{bmatrix} .45 & -.23 \\ .23 & .45 \end{bmatrix} \quad \mathbf{b}_1 = \begin{bmatrix} 0 \\ .28 \end{bmatrix} \quad p_1 = .2$$

$$f_2 : \quad A_2 = \begin{bmatrix} .52 & -.13 \\ .13 & .52 \end{bmatrix} \quad \mathbf{b}_2 = \begin{bmatrix} 0 \\ .40 \end{bmatrix} \quad p_2 = .2$$

$$f_3 : \quad A_3 = \begin{bmatrix} -.57 & .05 \\ .05 & .57 \end{bmatrix} \quad \mathbf{b}_3 = \begin{bmatrix} 0 \\ .40 \end{bmatrix} \quad p_3 = .2$$

$$f_4 : \quad A_4 = \begin{bmatrix} -.50 & .23 \\ .23 & .50 \end{bmatrix} \quad \mathbf{b}_4 = \begin{bmatrix} 0 \\ .32 \end{bmatrix} \quad p_4 = .2$$

$$f_5 : \quad A_5 = \begin{bmatrix} 0 & 0 \\ 0 & .45 \end{bmatrix} \quad \mathbf{b}_4 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad p_4 = .2$$

In Figure 8, we see that $S = f_1(S) \cup f_2(S) \cup f_3(S) \cup f_4(S) \cup f_5(S)$. However, it wouldn't be quite right to say that the “tree” S is the union of five “copies” of itself, because the ”trunk” $f_5(S)$ is just a line segment, not a little tree. In order to partially classify the different ways in which a set may be the union of “copies of itself,” terms like “self-similar” and “self-affine” have been introduced. We won't be too picky about the use of these terms, because the exact definitions are too restrictive for artistic (and even scientific) purposes. However, “self-similar” has a nice ring to it. It captures the essence of most fractals and many natural phenomena, so we will use it freely. For instance, it seems appropriate to call the set S in Figure 8 self-similar, despite the presence of the skinny set $f_5(S)$. It also seems fitting to describe the set S in Figure 5 as self-similar, even though the sets $f(S)$ and $g(S)$ overlap a bit (thereby disqualifying S as being self-similar, according to some technical definitions).

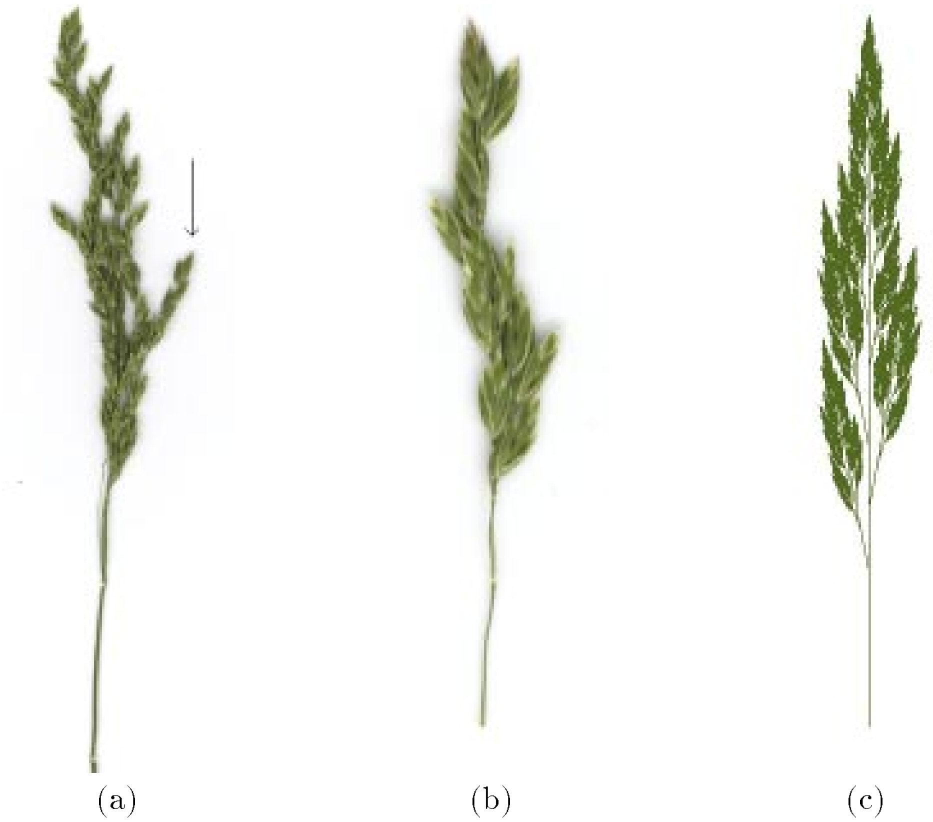

(a) (b) (c)

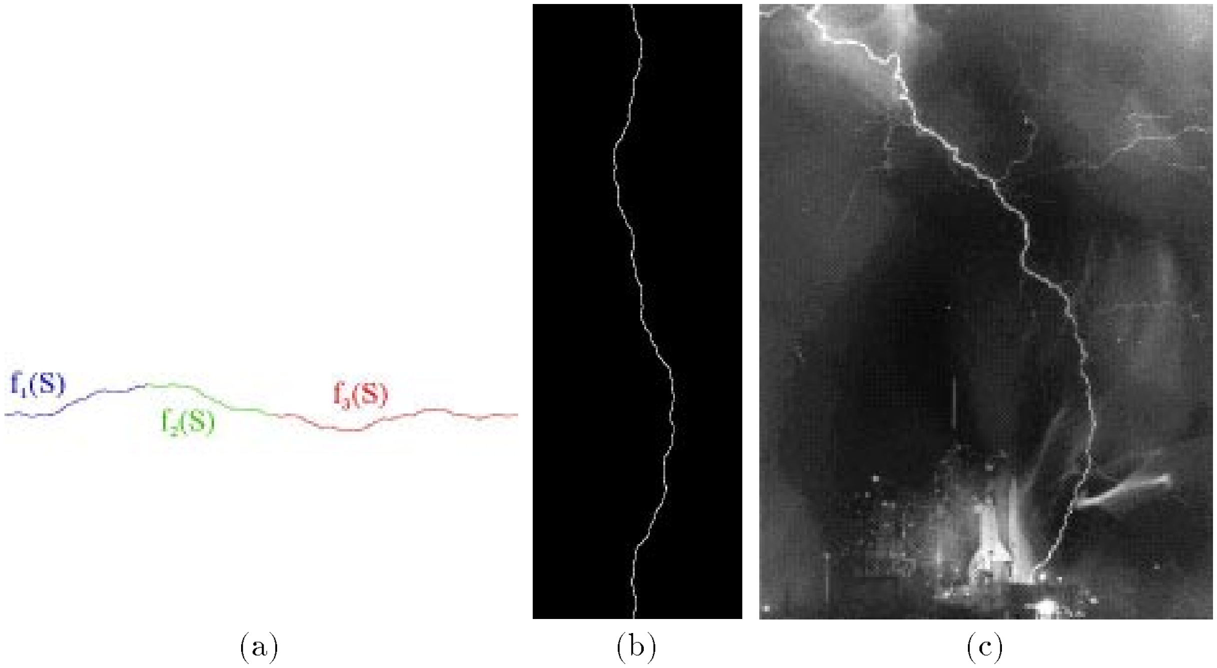
Figure 9. Self-similarity in grass: a part looks like the whole.

Another example of self-similarity in plants can be seen in Figure 9 (a) and (b). The picture in (a) was made by scanning a stem of grass in a flatbed scanner. A small sub-stem, indicated by an arrow, was broken off, then scanned at a higher resolution and magnified, resulting in (b). Notice how much the image in (b) resembles the whole stem in (a). Part of this resemblance is due to some fortuitous squashing in the scanner, but it's fairly obvious that some self-similarity is present. It therefore seems likely that the grass stem could be modeled reasonably well by fractal geometry. A first approximation (an artist would say a rough sketch) of the grass stem is the IFS attractor in Figure 9(c).

You shouldn't conclude from our discussion so far that the self-similarity of IFS attractors makes them suitable only for modeling plants. As far as natural forms go, there are many shapes in nature which are approximately self-similar. For instance, consider the IFS

$$\begin{array}{lllll}
f_1: & A_1 = \begin{bmatrix} .28 & .06 \\ .06 & -.28 \end{bmatrix} & \mathbf{b}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} & p_1 = .33 \\
f_2: & A_2 = \begin{bmatrix} .25 & .06 \\ -.06 & .25 \end{bmatrix} & \mathbf{b}_2 = \begin{bmatrix} .28 \\ .06 \end{bmatrix} & p_2 = .33 \\
f_3: & A_3 = \begin{bmatrix} .47 & 0 \\ 0 & -.47 \end{bmatrix} & \mathbf{b}_3 = \begin{bmatrix} .53 \\ 0 \end{bmatrix} & p_3 = .34
\end{array}$$

whose attractor S looks like Figure 10(a).



(a) (b) (c)

Figure 10. Fractal lightning and real lightning. (Photo courtesy of NASA.)

Jagged fractal curves like this make nice models of things like rivers, coastlines, profiles of mountain ranges or clouds, etc. For example, if we turn the attractor sideways (Figure 10(b)) it compares nicely with a lightning bolt (Figure 10(c)). Later, we'll learn a *quantitative* method for making such comparisons when we discuss a concept called *fractal dimension*. As another example, we can use the curve as the outline of a mountain range, as in Figure 11. In this image, we reversed the curve (reflected it in a vertical line) to make the profile of a second mountain range. The clouds in Figure 11 are also an IFS attractor which we'll discuss later. The mountains and clouds were put together in Adobe Photoshop, and some digital "airbrushing" was added to make the atmospheric haze.

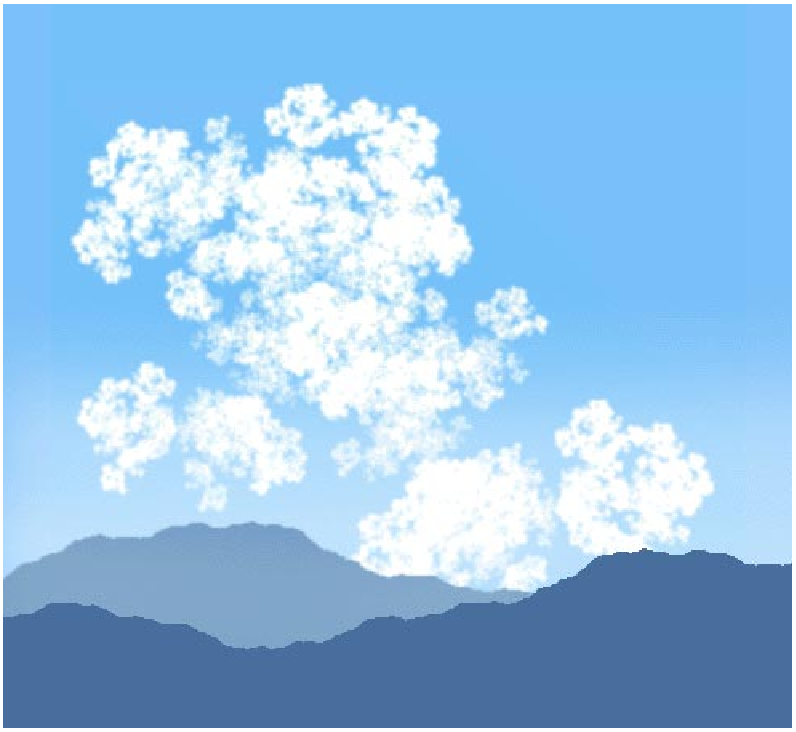


Figure 11. A landscape made from IFS attractors.