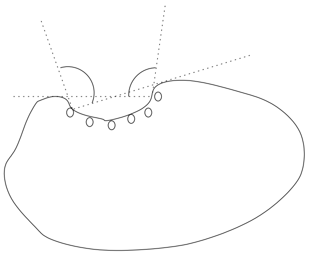
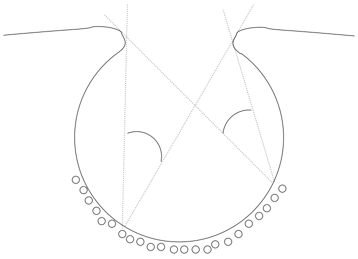
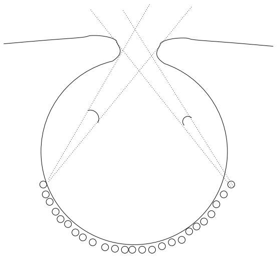
[**ASIDE:** These lecture notes do not replace the slides. Rather the two are complementary. I am not including all the figures from the slides. Instead, you should follow along the slides as you read.]

The origins of spatial vision

Our eyes are very sophisticated optical devices, and it took millions of years of evolution to reach this level of sophistication.[1] It is not known exactly how eyes evolved, but it is believed that the earliest eyes consisted of a small number of light sensitive cells distributed over a small region on the outer surface of an animal (and hooked up to a primitive nervous system). Let's suppose the cells are distributed over a concave pit such as in the figure below. Six light sensitive cells are shown. Because the pit is concave, each cell will receive light from a limited set of directions. For the leftmost and rightmost cells, the range of directions of light coming from the scene is shown in the figure on the left. (See slides for updated figures.)







Now suppose that something to the left of this animal were to move towards the animal and block the partially block the sky (casting a shadow). The leftmost cell only receives light from above/right and so the light received by the cell would not be affected. (See sketch in lecture slides). The rightmost cell receives light from above/left and so the approaching animal would block the skylight and the rightmost cell would received less light. The changing measurement would tell the animal that something dark is now presen on the left. A defensive response of the animal therefore might be to move toward the right, i.e. away from the approaching animal. (Alternatively, the animal could move to the left, which would be a more aggressive response.) Such a response might allow the animal to survive. If it produced offspring with similar concave light-sensitive regions which produced similar actions to changing light measurements, they might also have a better chance of surviving.

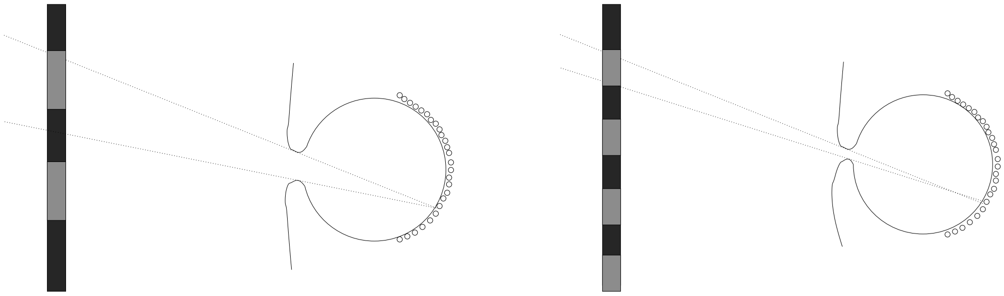
One way to improve this vision system would be to make the eye more cavelike, as in the two figures above to the right. Here we are reducing the *aperture* by which light can enter the concavity. This reduces the angle of incident light that reaches each photoreceptor cell of the eye.

The advantage of reducing the aperture is that each cell receives light from a more restricted set of directions which provides the eye with more detailed information about the directional distribution of light arriving at the aperture. The understand this, consider the figures below. The pattern of light in the scene is an alternating dark grey and light grey, and each cell averages over some dark regions and some light regions.

[1] See interview with Richard Dawkins (outspoken evolutionary biologist at U. Oxford) `https://www.youtube.com/watch?v=bwX3fxOZg5o`

In the figure on the left, the aperture is so big that each cell would see part of a light grey region and part of a dark grey region. If the aperture were slightly larger, so that each cell received light from an equal amount of light and dark regions, then each cell would receive the same total amount of light. In this case, we would say that the light and dark greys had been *blurred* away completely.

In the figure on the right, the aperture has been decreased, and the pattern of light and dark grey has a higher frequency. This figure is drawn such that the angular width of the aperture is exactly matched to the width of the pattern on the surface. Notice that the retinal image would still be blurred, since most cells would receive a mix of light from light and dark regions and only a few cells would see only a light gray or only a dark grey. The retinal image would look more like a sinusoid with smooth transition from dark to light, rather than the piecewise constant intensity (light,dark,light,..) in the scene.



The disadvantage of using a smaller aperture is that it reduces the amount of light reaching each cell. Eventually, if the aperture becomes a "pinhole" and the image will be very dark indeed. Next lecture I will discuss how lenses avoid this problem.

Units of angle

We will talk about angles in various ways today. As you know from Calculus, it is common to define an angle in units of degrees or radians. Recall that 2π radians is 360 degrees, or

$$\frac{360 \text{ degrees}}{2\pi \text{ radians}} = \frac{180 \text{ degrees}}{\pi \text{ radians}} \approx 57 \text{ degrees per radian.}$$

When doing vision calculations, it is common to make a small angle approximation.

$$\theta \approx 2\tan(\frac{\theta}{2}).$$

This approximation essentially says that the length of an short arc of a circle is approximately equal to the length of the line segment joining the end points of that arc of a circle.

In astromony, it is necessary to refer to angles that are much smaller than a degree. In particular, there are 60 *minutes* in one degree, and there are 60 *seconds* in one minute. Minutes of arc comes up often in vision. Rarely do we need to talk about seconds, but it does come up.

Aperture angle and f-number

Let's returning to our discussion of apertures. The notion of an aperture should be familiar to those of you who dabble in photography. Let's consider a camera rather than an eye. The camera is a hollow box and light enters the box through a hole.

Ignore the lens for today and just think about the hole or aperture. Let A be the diameter of the aperture and let f be the distance from the center of the aperture to the sensor surface. Then, if $A \ll f$ which is usually the case, then we can make a small angle approximation, namely A/f is approximately the angle subtended by the aperture as seen from a point on the sensor on the image plane at the back of the camera box.

The amount of light reaching a point on the image plane depends on what is visible in the 3D scene in directions "seen" by that pixel. It also depends on the angular size of the cone of light rays that reaches this image point. To think about angles, it is best for now to just consider a 2D scene rather than 3D scene. So, the camera is a square and the aperture is gap in the square and the sensor is a line. (See figure in slides.) The angle subtended by the aperture is approximately $\frac{A}{F}$ radians, where A is the width of the aperture and F is the distance from the aperture to the sensor. Its inverse, $\frac{F}{A}$ is called the *F-number.*

If you have done any photography using an SLR camera, then you are familiar with f-number, as it is one of the main parameters you can manipulate. When you change the f-number, in fact you are just changing the aperture since f is fixed. The effect of course is to change the amount of light that reaches the lens, making the image brighter or darker. There are other effects as well, as we'll see once we consider lenses.

What are some typical f-numbers? Camera's often have f-numbers[2] ranging from about 2 to 16. For example, if $A = 5mm$ and $F = 50mm$, then the f-number $\frac{F}{A}$ is 10. A small angle approximation works quite well here. We can also define an f-number for a human eye. Typical values of the aperture (pupil) diameter are $A = 5mm$ and $F = 25mm$, for an f-number of 5. A small angle approximation still holds for these values too.[3]

Visual angle

A second fundamental angular quantity is the angle subtended by an object in the 3D world as seen from a position in space. This angle is called the *visual angle* subtended by this object. Assuming that the object's height (or width) is small compared to the distance to the object, we can make a small angle approximation and define:

$$\text{visual angle (radians)} = \frac{\text{height of object}}{\text{distance to object}}$$

Let's suppose that the aperture angle is very small (large f-number) and treat the aperture as a point in space. This is usually called a *pinhole camera.* From high school geometry reasoning – namely, opposite angles are equal – we know that the visual angle subtended by the object can be

[2] Technically it is the camera and the lens together that define the f-number since the aperture is defined in the lens body.

[3] You may be asking yourself how you know when a small angle approximation is "good enough". There is no single answer to this. It really depends on the precision you need.

written equivalently as:

$$\text{visual angle (radians)} = \frac{\text{height of image (of object) on sensor}}{\text{distance from pinhole to sensor}}$$

For example, consider your thumbnail which is about 1 cm wide. Suppose you view your thumb at an arm's length distance say about $57 = \frac{180}{\pi}$ cm. The thumbnail would have a visual angle of $\frac{1}{\frac{180}{\pi}} = \frac{\pi}{180}$ radians. Converting to degrees by multiplying by $\frac{180}{\pi}$ degrees/radian gives us 1 degree, i.e. a thumbnail at arm's length subtends about 1 degree of visual angle.

Here is a second example. Consider a person's head a large distance, say 18 m. Suppose the person's head is about 30 cm high. To make the calculation easier, say it is 31.4 cm high, or $\frac{\pi}{10}$ m. Then the visual angle subtended by this person's head would be about $\frac{\pi/10}{18}$. Converting to degrees by multiplying by $\frac{180}{\pi}$ gives 1 degree, i.e. the visual angle of the person's head at that distance would be 1 degree.

A third example of a visual angle is the moon which subtends about half a degree, or 30 minutes (arcmin). You may have heard of the moon illusion, which is that the moon appears much bigger when it is near the horizon than when it is overhead. This is not an optical effect due to bending of light through the atmosphere, as some people assume. Rather, it is a perceptual effect. This illusion is very strong and has been studied in great detail literally for centuries. (Read the wikipedia article if interested.)

Image position

We would like to define positions of points in an image and relate these image positions to positions in the 3D scene. Define a coordinate system with axes XYZ such that the origin $(X, Y, Z) = (0, 0, 0)$ is at the center of the camera/eye aperture. We'll just ignore the aperture itself for the rest of today and assume a pinhole camera. Let $(\hat{X}, \hat{Y}, \hat{Z})$ be the coordinate axes. Let the Z be depth variable and XY be the axes parallel to the image plane. Typically X is the right right and Y is up. The $\hat{Z}$ axis is called the *optical axis*.

Let's next relate positions XYZ in the 3D scene with positions on the image plane. We begin by reviewing some of the basic geometry of image formation. Consider a 3D scene point (X_0, Y_0, Z_0) in this coordinate system. Suppose that the image plane is behind the camera at a distance f. The line through this scene point and through the origin (pinhole) intersects the *image plane* $Z = -f$ at position $(x, y, -f)$. The point of intersection is the *image position.* So, what we have just done is project the 3D point onto the image plane.

Using high school geometry (similar triangles), we can see that

$$\frac{x}{f} = \frac{X_0}{Z_0}, \qquad \frac{y}{f} = \frac{Y_0}{Z_0}$$

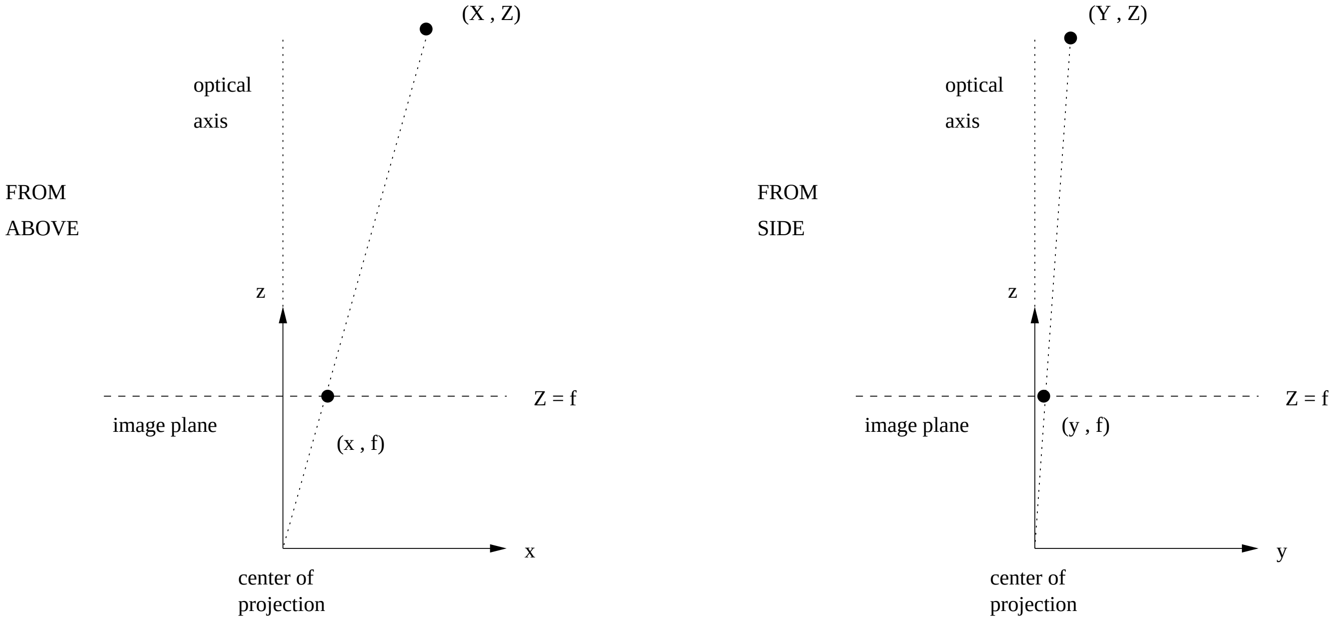
Note that if (x, y) is close to the center of the image, then we can make a small angle approximation and talk about $\frac{x}{f}$ and $\frac{y}{f}$ as angles (in radians).

Notice that the image will be upside down and backwards, in the sense that if the object points has $X_0 > 0$ then $x < 0$ and if $Y_0 > 0$ then $y < 0$. It can be confusing to think of images that are upside down and backwards. Typically instead one thinks of the images as upright, as we discuss next.

Visual direction

Consider a plane at $Z = f$ in front of the eye or camera. This is not a real image plane, but it is still useful to think about. Just as we projected the 3D scene though a pinhole and onto an image plane behind the camera, we can project the 3D scene towards a pinhole and consider where the image point intersects the plane $Z = f$ in front of the camera.

$$(x, y) \ = \ f(\frac{X_0}{Z_0}, \frac{Y_0}{Z_0}). \tag{1}$$



Such a point defines a visual direction from the eye/camera out to the scene. With these (x, y) coordinates, if $X > 0$ then $x > 0$ and $Y > 0$ then $y > 0$. We will typically use this coordinate system (x means to the right, y means up). Moreover, we are typically interested in visual direction rather than some position in an arbitrary plane $Z = f$. We will refer to visual direction $(\frac{x}{f}, \frac{y}{f})$. If these values are small, then they are approximately angles in radians.

Depth map

For every image position (x, y, f) in this abstract image plane in front of the camera, there is typically one surface point that is visible along the ray from the center of projection through that position. The function $Z(x, y)$ maps each position in the (x, y) projection plane to a depth in the world. This function is called the *depth map*.[4]

Notice that the *depth* is not the Euclidian distance $\sqrt{X^2 + Y_2 + Z^2}$ to the 3D point (X, Y, Z). Rather we are only considering the Z value. If we are looking a wall that directly in front of us, then all points on the wall would have the same depth, even though the Euclidian distance would vary along the wall.

[4] Note that we could alternatively define the depth map to be a function of visual direction $Z(\frac{x}{f}, \frac{y}{f})$.

Example: Ground plane

Consider a specific example of a depth map. Suppose the only visible surface is the ground, which we approximate as a plane. Suppose the camera/eye is height h above this *ground plane*. That is, the ground plane is

$$Y = h$$

where $h < 0$. We are still assuming the camera is pointing in the Z direction.

What is the depth map of the ground plane? From Eq. (1), we substitute $-h$ for Y:

$$(x, y) \; = \; f(\frac{X_0}{Z}, \frac{h}{Z}).$$

In particular,

$$Z(x, y) \; = \; \frac{-hf}{y} \tag{2}$$

Thus, the depth map $Z(x, y)$ does not depend on x. It only depends on y. When $y = 0$, we have $Z = \infty$. This is the *horizon*. When $y < 0$, we have $Z > 0$. (Note that $f > 0$.) These are the visible points on the ground and we see that closer points to the eye (smaller Z) have more negative y. What about points where $y > 0$. These are not points on the ground, and their depths are not defined in Eq. (2). If there is nothing in the scene other than a ground plane, then the points where $y > 0$ would be the sky. We could take the depths at $y > 0$ to be infinity.

Note that the depth map for a groud plane only depends on y, and not on x. For any fixed y, all visible points along that horizontal image line have the same depth (independent of x). Also, points of a fixed depth $Z = Z_0$ all project to the same y value. Again, $\frac{y}{f}$ is the angle of a point below the visual horizon. This angle varies inversely with depth.

Binocular disparity

Having two eyes gives us two slightly different views of the world, and the slight differences provide information about depth. Let's begin by assuming that we have two eyes or cameras and that the optical axes of the two eyes are parallel, i.e. the eyes have the same Z direction. Let the right eye be positioned at point $(0, 0, T_X)$ in the left eye's coordinate system. The distance T_X is sometimes called the *interocular distance*. With these assumptions, a 3D point with coordinates (X_0, Y_0, Z_0) in the left eye's coordinate system would have coordinates $(X_0 - T_x, Y_0, Z_0)$ in the right eye's coordinate system. As such, this 3D point would project to a different x value in the left and right images. The difference in x position of that point is called the *binocular disparity*. In human vision, it is more common to define it in terms of visual direction, so that's what we will do.

$$\text{disparity (radians)} \; \equiv \; \frac{x_l}{f} - \frac{x_r}{f} \; = \; \frac{X}{Z_0} - \frac{X - T_x}{Z_0} \; = \; \frac{T_x}{Z_0}.$$

Note that we are assuming the eyes are parallel and pointing forward. In this case, the y values of the projected points are the same in the two eye images.

Vergence

To visually explore the world around us, we rotate our eyes namely we point the optical axes of each eye at a particular position in 3D space. The 3D point that we "look at" is typically not at infinity, and so the two eyes and the 3D point form a triangle. We say that the eyes *converge* on this 3D point, and the angle of the triangle at the 3D point the eyes are looking at is the *vergence angle.*

Rotating (verging) the eyes changes the binocular disparity. If we look at a point with the two eyes, it means that we are rotating the eyes such that we set (x_l, y_l) and (x_r, y_r) to $(0,0)$. In particular, the 3D point that the eyes are looking at will have zero disparity. Other points will have positive or negative disparity depending on whether these points are in front of or behind the 3D point we are looking at. We will discuss this in more detail in a few weeks. For now let's sketch out just the basics of how this works. See the accompanying slides which show a sketch of two people, a ground plane, and a horizon.

Let the left and right eyes rotate to the left or right by angles θ_l and θ_r radians. Note that we assume the rotation is left and right, i.e. about the Y axis. Suppose that this 3D point had horizontal angular directions x_l and x_r in the left and right eye when the eye's were pointing straight ahead. Since the eyes have been rotated by θ_l and θ_r respectively, these rotations bring the binary disparity of this 3D point to $\frac{x_l}{f} - \theta_l$ and $\frac{x_r}{f} - \theta_r$ in the two eyes, respectively. This would change the disparity to

$$\begin{aligned} \text{disparity (radians)} &= (\frac{x_l}{f} - \theta_l) - (\frac{x_r}{f} - \theta_r) \\ &= (\frac{x_l}{f} - \frac{x_r}{f}) - (\theta_l - \theta_r). \end{aligned}$$

Thus, the effect of rotating the eyes horizontally (left/right) is to shift *all* the points in each image by angles θ_l and θ_r in the left and right eyes respectively. This changes the disparity of all points by a constant $\theta_l - \theta_r$.