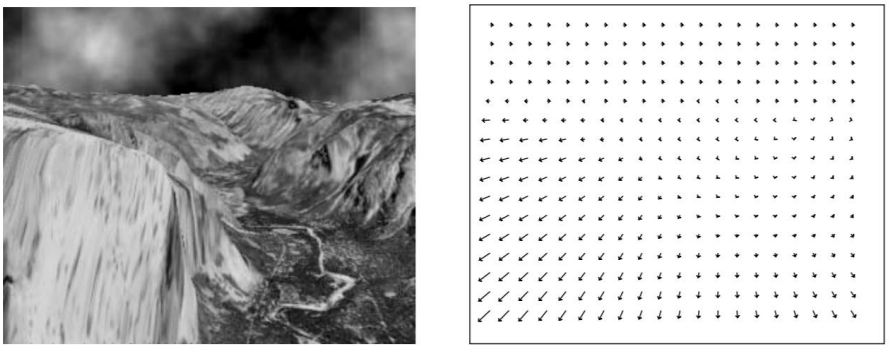
Motion Field

In lecture 8, we examined the computational problem of estimating the motion (v_x, v_y) at a point (x, y) in the visual field. The idea was to measure local derivatives of image intensity, and to use these derivatives to constrain the possible velocity vectors. The main assumption was that moving points do not change their intensity over time, and indeed that was the defining property of a moving point.

Today we are going to consider not just one point, but all the points (x, y) in an image. Let's say the scene depth map is $Z(x, y)$ and we would like to know the image velocity (v_x, v_y) for each (x, y). We will assume that the image motion is due to motion of the eye/camera, and that the scene itself is static. In this case we can write down simple formulas for how the velocity (v_x, v_y) at each point in the image depends on the motion of the observer and on the depths of the scene points. These velocities define the instantaneous *motion field*.

As an example, consider a single frame from a video known as the *Yosemite sequence*. This was a computer graphics generated video of a fly through through the Yosemite Valley in California[1] Because was computer generated, it had a well defined depth map $Z(x, y)$ and one could compute a vector field — (v_x, v_y) at each pixel – shown on the right.

Today we will look at the motion fields that arise from different observer motions and different scene layouts.



Translation of viewer

I'll first discuss observer motions that consist of a change in observer position, but no observer rotation. Suppose that the viewer changes position over time by moving in a straight line over a short time interval, and does not rotate during this motion. Because the viewer observes the scene from different positions, the projected positions of objects in the image change too.

Suppose the camera translates with 3D velocity (T_x, T_y, T_z). For example, forward camera motion with unit speed is 3D velocity $(0, 0, 1)$. Rightward camera motion with unit speed is 3D velocity $(1, 0, 0)$. Upward camera motion is $(0, 1, 0)$. When the camera translates, the position of any visible point varies over time. In the camera's coordinate system, the position of the point moves with a velocity vector opposite to the camera. If the camera coordinates of a point at time

[1] It was often used in early computer vision research (1980's and 1990's) to test the accuracy of computer vision methods for estimating image motion.

$t = 0$ are (X_0, Y_0, Z_0), then at time t the point will be at $(X_0 - T_x t,\ Y_0 - T_y\ t,\ Z_0 - T_z\ t)$ in camera coordinates.

Now let's project the 3D point into the image plane. How does the image position of this point in the image vary with time? We will use a visual field projection plane $Z = f$ in front of the viewer. *We would like to expression image position in terms of visual directions, and so we set* $f = 1$. *Thus, x and y are in units of radians.* Note this requires a small angle approximation.

The image coordinate of the projected 3D point is a function of t, namely,

$$(x(t), y(t)) \ = \ (\frac{X_0 - T_x\ t}{Z_0 - T_z\ t}, \frac{Y_0 - T_y\ t}{Z_0 - T_z\ t})$$

Taking the derivative with respect to t at $t = 0$ yields an *image velocity vector* (v_x, v_y) in radians per second:

$$(v_x, v_y) = \frac{d}{dt}(x(t), y(t))\ |_{t=0} \ = \ \frac{1}{{Z_0}^2}(-\ T_x\ Z_0\ +\ T_z\ X_0, -\ T_y\ Z_0\ +\ T_z\ Y_0). \tag{1}$$

The velocity field depends on image position (x, y) and on the depth Z_0 and on (T_X, T_Y, T_Z). We next decompose the velocity field into a lateral component and a forward component.

Lateral component of translation

Consider the case that $T_z = 0$. This means the viewer is moving in a direction perpendicular to the optical axis. One often refers to this as *lateral motion*. It could be left/right motion, or up/down motion, or some combination of the two. Plugging $T_z = 0$ into the above equation yields:

$$(v_x, v_y) = \frac{1}{Z_0}(-T_x, \ -T_y) \ .$$

Note that the direction of the image velocity is the same for all points, and the magnitude (speed) depends on inverse depth.

A specific example is the case $T_y = T_z = 0$ and $T_x \neq 0$. The motion field corresponds to an observer looking out the side window of a car, as the car drives forward. Another example is the case that the scene is a single ground plane: recall the relation $Z = \frac{h}{y}$ from lecture 1. The image velocity is then

$$(v_x, v_y) = -\frac{T_x}{h}(y, 0).$$

The minus sign is there because the image motion is in a direction opposite to the camera motion. The speed is proportional to y is a result of the depth of the ground plane being inversely proportional to y, e.g. the depth is ∞ for $y = 0$ which is the horizon. *See the examples given in the slides.*

Lateral motion is very important for vision. Our eye position continuously shifts over time. If when we think we are still, in fact we are continously shifting our weight and changing our pose. This is in part to relieve our joints and muscles, but it also provides visual information for maintaining our pose. As we lean to the left, the visual scene drifts slightly to the right, and vice-versa. We rely on this motion field the stabilize ourselves with respect to the surrounding world.

This reliance of the motion field becomes evident when we stand in front of a cliff, so that the ground in front of us is tens or hundreds of metres away. Normally, the ground in front of us moves

opposite to us as we sway slightly back and forth. But when we stand in front of a cliff, there is essentially no lateral motion (visual) field because Z is so big and $\frac{1}{Z}$ is near 0. This lack of motion is problematic for visually controlling our posture. It is the main reason we get dizzy (vertigo) when we stand at the edge of a cliff. More generally, it is one of the factors that contribute to a fear of heights. It is also why it is more difficult to do fancy balance poses in yoga when you are looking up at a high ceiling than when you are looking down at the ground in front of you. (A low ceiling works fine.)

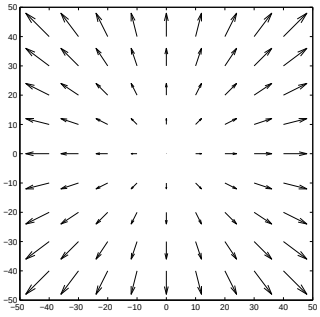
Forward translation

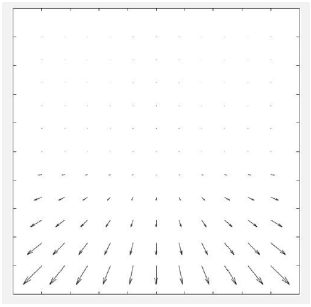
In case of forward translation ($T_x = T_y = 0$ but $T_z > 0$), Eq. (1) becomes

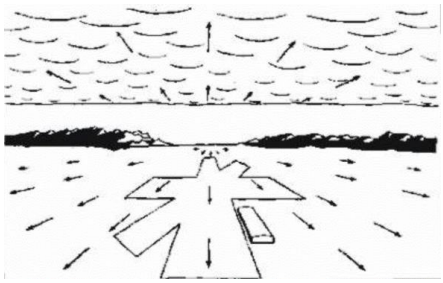
$$(v_x, v_y) \;=\; \frac{T_z}{Z_0}\,(x, y). \tag{2}$$

By inspection, this field radiates away from the origin $(x, y) = (0, 0)$. Also, the speed (i.e. the length of the velocity vector) is :

- proportional to the angular distance $\sqrt{x^2 + y^2}$ from the origin
- inversely proportional to the depth Z_0
- proportional to the forward speed of the camera T_z.







The example on the left is for a wall (constant depth). The middle panel shows the case of a ground plane, which has depth map $Z = h\frac{f}{y}$ and so:

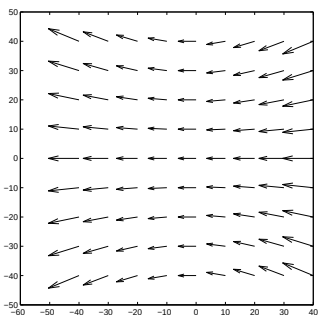
$$(v_x, v_y) \;=\; = \frac{T_z}{h}\,(\,xy\,,\;y^2)$$

Note that in this case the velocities near the horizon $y = 0$ are small. This is a familiar case of walking forward. Another situation in which this arises is what a pilot sees when landing a plane. This scenario was one of the first applications in which psychologists studied this 'direction of heading' problem. (The illustration on the right above is taken from a classic book by J. J. Gibson in 1950.)

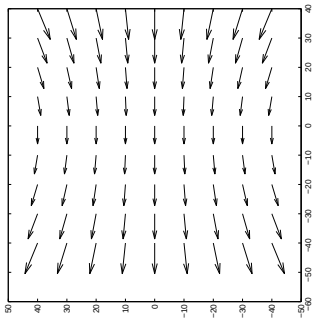
Rotation of viewer

The viewer can not only change position over time. It can also change the direction of gaze over time. This can be done by rotating the head[2] or by rotating the eyes within the head, or both.

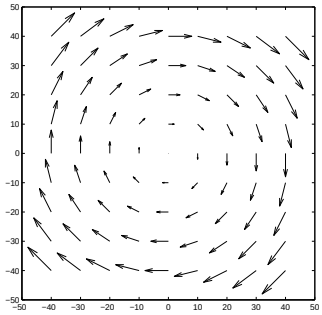
Rotation induces a smooth motion field on the retina. Basically, pan is rotation about the Y axis $\Omega = (0, 1, 0)$ and produces roughly motion in the x direction. Tilt is rotation about the X axis $\Omega = (1, 0, 0)$ and produces roughly motion in the y direction. Roll is rotation about the Z axis $\Omega = (0, 0, 1)$ and produces circular motion where the speed depends on distance from the center. One can derive exact expressions for the velocity fields by projecting onto an image plane $Z = 1$ as we did for translation. The resulting expressions are a bit more complicated than in the translation case (note the non-linearities in the pictures below), so I will skip them and just show the pictures.



$\boldsymbol{\Omega} = (0, 1, 0)$
tilt



$\boldsymbol{\Omega} = (1, 0, 0)$
pan



$\boldsymbol{\Omega} = (0, 0, 1)$
roll

Eye Movements

Let's next consider several different types of rotation in human vision: smooth pursuit eye movements, the vestibulo-ocular reflex, and saccades. These types of motion have different roles and are computed by different parts of the visual system.

A few basic facts before we begin: Eye rotations are controlled by muscles that are attached to the side of the eyeball. See the figure in the slides. There is a pair of opposing muscles for each of the three rotation directions. These muscles are signalled directly by motor (output) neurons whose cell bodies are in the midbrain. The axons from these motor neurons are bundled together into the *oculomotor nerve*, which carries carry signals to the muscles that rotate the eye, as well as blink, accommodation, and pupil contraction commands.

Smooth pursuit eye movements

Smooth pursuit eye movements are voluntary eye rotations that keep a desired object stable in the center of the visual field (fovea). An example is the eye movements that you make when you visually track something moving the world e.g. when you watch a cyclist ride by, or when you are moving and you look at some object that is fixed in position. These eye movements are relatively slow, since you and the objects that you track tend to move relatively slowly in the world. There are

[2]Note that when you rotate your head with the eye fixed in head coordinates, it induces both a translation and a rotation of the eye, since the head rotates around some point in the neck. Lets only concern ourselves with pure rotation for the moment.

limits on how fast object can move relative to you (in terms of visual angle per second) for you to be able to track it. For example, if I move my finger in front of your eye, you can track it but only up to some limited speed and only if the motion is smooth enough to be predictable.

The speed limitation for the smooth pursuit system is that you need to compute image velocity of the point(s) you are tracking and computing image velocity takes some time – since the signal needs to travel to V1 and then MT and beyond. The brain also needs to compute the rotation that is required to correct for the slippage – i.e. reduce the image motion to 0 – and it needs to compute motor commands to rotate the eye to reduce the image motion to 0. These motor commands need to be sent to the midbrain where a signal can be sent on the oculomotor nerve to the muscles that control the direction of the eye. (The various neural pathways are well known, but I am omitting the details here since I just want to make a general point about why the system is relatively slow.)

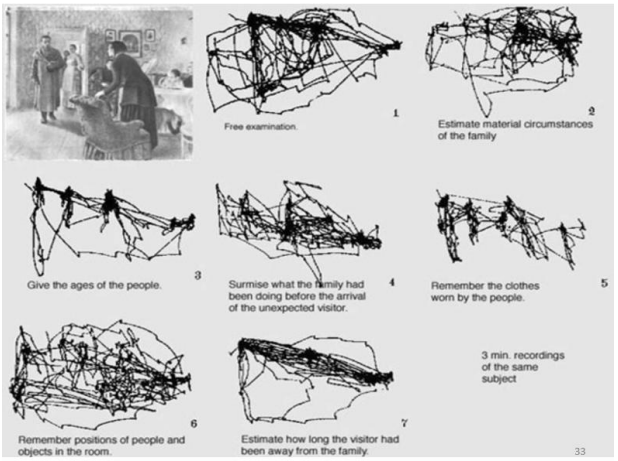
Vestibulo-ocular reflex (VOR)

The vestibulo-ocular reflex is another type of eye movement that is used to reduce image motion on the retina to zero. However, it is quite different from smooth pursuit. In general when the head moves – whether it is translation or rotation or both – the motion causes a shift in the retinal image, as we discussed earlier in the lecture when we described egomotion. However, sometimes the head motion is just due to postural adjustment or a turning of the neck. (Think for yourself how difficult it is is to *not* make such motions. It would require that you are rigid as a statue!) The role of VOR is to quickly sense this head motion and to rotate the eyes to compensate for it and to keep the retinal image as stable as it can. For example, fixate at one of the words on this page (or screen) and then rotate or translate your head left and right and remain fixated on that word. You will find this is very easy to do and indeed it is automatic. This is VOR.

VOR depends on the vestibular system which is part of your inner ear. The vestibular system senses linear and rotational acceleration of the head. There are two parts – see slides. The first part detects rotational acceleration. It consists of three loops called the *semi-circular canals*. The semicircular canals are filled with fluid, and when the head rotates, the fluid moves in the canal and this fluid motion is sensed by mechanical receptors. (These are analagous to photoreceptors, except that they sense fluid motion rather than light.) These receptors encode the fluid motion, and the code is sent from the ear along the auditory nerve to the brain.

[ASIDE: If you stand up and turn around and around several times, then as your head rotates, the fluid drags along and eventually it will have the same speed as the canal itself. Then when you stop rotating, the fluid will keep going and again the system will senses the fluid motion relative to the canal, which sends a (erroneous) signal that the head is rotating again. And you will fall down. I'm sure you all did this when you were children.]

The second part of your vestibular system measures linear acceleration. How does this work, intuitively? Imagine a grassy surface with stones sitting on it. If the surface is suddenly moved sideways, then the stones will roll relative to the surface and the grass under the stone will be affected by this motion. The "grass" is a set of mechnical receptors and the stones are just that – small stones (called otoliths). This analogy allows us to think about linear accelertion in two directions, namely parallel to the surface. To sense linear acceleration in three dimensions, you need surfaces with more than one orientation. (There are two such surfaces, called the saccule and utricle.)



The VOR is extremely fast, and the reason this is possible is that the circuit is short, namely the signal is sensed in the inner ear, and then the signal goes from there to the brainstem, from which signals are sent to the eye muscles driving compensatory motion.

VOR does not depend on a visual signal, and indeed works even when the eyes are closed. You can verify this for yourself. Look at some object in the scene, and close your eyes. Now shake your head back and forth and keep trying to fixate the imagined location of the object. Your eyes will rotate as you do so, but will keep fixation (within say 5 deg of visual angle) on whatever you had been looking at before you closed your eyes.

[ASIDE: the vestibular system doesn't measure the rotation of head directly, but rather it measures changes in rotational velocity over time (or rotational acceleration), and it doesn't measure the translation (T_X, T_Y, T_Z) directly but rather it measures the change $\frac{d}{dt}$ in the translation velocity over time. The system needs to integrate the changes in rotation or translation over time in order to maintain an estimate of the rotation velocities or the translational velocities themselves.]

Saccades

The last type of eye movement that we consider is the brief and rapid jumps that you make when scanning a scene. These are called *saccades* and they occur typically several times per second.

As an example of saccades, consider the classical experiment by Yarbus from the 1960's which was the first to tie eye movements to image content. Yarbus measured the direction in which a person was looking as the person examined a painting (called *The Unexpected Visitor*). He asked the person to answer different questions about the painting and he showed that the eye movement pattern depended on the question that the person was asked. (This is not so surprising, and I would say the main contribution here was to introduce a new experiental method which has been used by many people since.)

Eye movement research is very active to this day, in part because of new methods that allow researchers to measure eye movements while people perform tasks in the world. It is possible now to mount cameras on people's heads and simultaneously monitor their eye movements while filming the scene the the person is looking at (by having cameras mounted to the head that point out into

the scene and other cameras that point at the eyes and measure their movement). The slides give links to videos that show the changing gaze direction superimposed on the scene. Note in the tea making video, the position of the fovea in the scene is indicated by a small white spot of light, and the video also shows the eye itself as it moves.

You are generally not aware that your eyes are moving around so much as you perform a task. As you read the words on the page, it feels like the image page is somehow stable and your eyes are moving around on the page ? But of course, that is not what's happening. The stable page is a construction of your mind/brain. In fact, the images that reach your retina are a sequence of *non-aligned* maps – which are each snapshots. The brain needs to stitch these maps together for you to have that feeling that you are looking at one unified map of the world around you. It is not well understood how this stitching together (and memory of the different snapshots is stored), and where this occurs in the brain. It is also unclear two what extent this stitching does indeed occur and to what extent these maps are just an illusion.

That's it for today. I hope you appreciate better now how challenging is the task we perform of moving around in the world and fixating and maintaining our gaze on the objects of interest. This is difficult enough for basic tasks like cooking food (making tea) or driving a car. It is all the more impressive to think of the level with which professional atheletes (such as hockey players) can perform these tasks.

