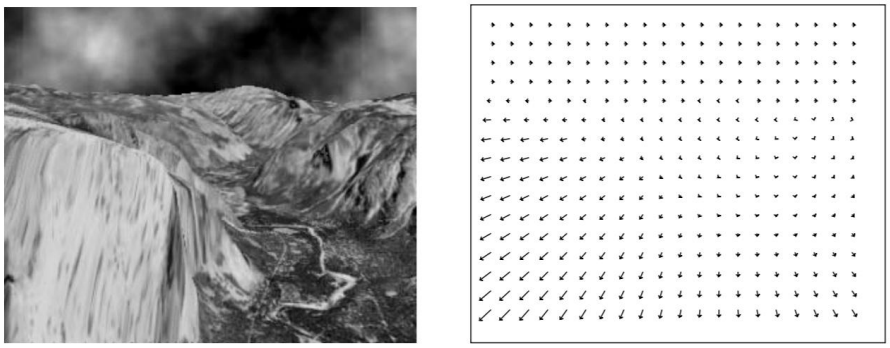
Motion Field

In lecture 7, 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. Lets say we have a depth map $Z(x, y)$ and we would like to know the image velocity (v_x, v_y) for each (x, y). We will assume that the motion is due to movement of the eye/camera. For simplicity, we will assume 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. I'll first discuss observer motions that consist of a change in observer position, but no observer rotation.



Translation of viewer

We 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* and express the position in radians. The image coordinate of the projected 3D point is a function of t, namely,

$$\frac{1}{f}(x(t), y(t)) \ = \ (\frac{X_0 - T_x\ t}{Z_0 - T_z\ t}, \frac{Y_0 - T_y\ t}{Z_0 - T_z\ t})$$

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

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

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. In 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 9, 10 which show two frontoparallel surfaces and a ground plane, respectively.

Lateral motion is very important for vision. Our eye position almost always 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 the sky or a high ceiling than when you are looking down at the ground in front of you.

Forward translation

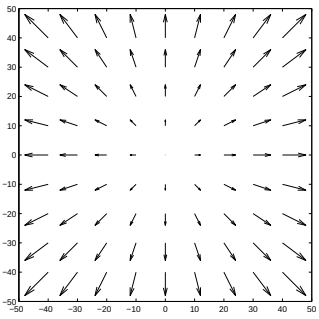
In case of forward translation ($T_x = T_y = 0$ but $T_z > 0$), Eq. (1) becomes

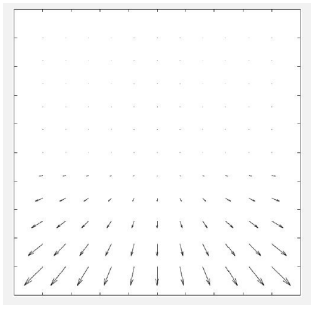
$$(v_x, v_y) \;=\; \frac{T_z}{Z_0}\,(\frac{x}{f}\;,\;\frac{y}{f}). \tag{2}$$

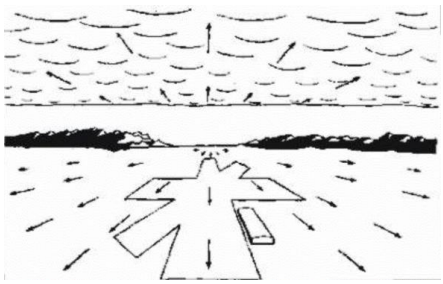
By inspection, this field radiates away from the origin $(x, y) = (0, 0)$. Also, the speed (i.e. the length of the velocity vector) is :

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

See the example on the left.







The middle panel shows the case of a ground plane, which has depth map $Z = h\frac{f}{y}$ and so:

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

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.)

General (non-lateral) translation

In the case that we do not have pure lateral translation, i.e. if $T_z \neq 0$, we can write the motion field slightly differently. Putting the lateral and forward components of the motion field together, we get

$$(v_x, v_y) = \frac{1}{Z_0}(-T_x, \ -T_y) \ + \frac{T_z}{Z_0}\ (\frac{x}{f}\ ,\ \frac{y}{f}) \tag{3}$$

$$= \frac{T_z}{Z_0}(-\frac{T_x}{T_z}, \ -\frac{T_y}{T_z}) \ + \frac{T_z}{Z_0}\ (\frac{x}{f}\ ,\ \frac{y}{f}) \tag{4}$$

$$= \frac{T_z}{Z_0}\ (\frac{x}{f} \ -\frac{T_x}{T_z}, \ \frac{y}{f} \ -\frac{T_y}{T_z}) \tag{5}$$

Define the special image direction:

$$(\frac{x_0}{f}, \frac{y_0}{f}) \ = \ \frac{1}{T_z}(T_x, T_y) \tag{6}$$

which is called the *heading direction*. Then,

$$(v_x, v_y) = \frac{T_z}{Z(x,y)}(\frac{x - x_0}{f}, \frac{y - y_0}{f}) \ .$$

Notice that the translation field diverges away from the heading direction. See example in slides.

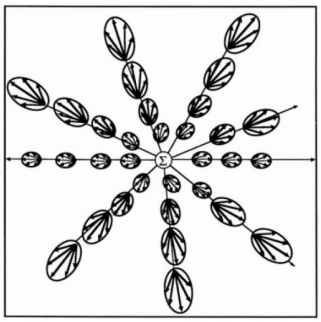
Computing the heading direction (from MT to MST)

How can a visual system estimate the direction in which it is heading? There are basically two steps. First, estimate the local velocities at as many points (x, y) as possible. Second, estimate a direction from which all velocity vectors point away.

As I discussed last lecture, the brain carries out the first step itself in two parts. (Cells in V1 measure normal velocity components, and cells in MT combines these normal velocity estimates to estimate velocities.) How does one brain compute the heading direction from these velocity estimates? This computation occurs in another area of the brain, known as MST which is close to MT. MST stands for "medial superior temporal". "Medial means inside (as opposed to lateral). Superior means on top. "Temporal" refers to temporal lobe.

Cells in MST receive direct inputs from cells in MT. MST cells have very large receptive fields. Many of these cells are sensitive to expanding patterns within their receptive field. You can think of these cells as getting excitatory input from MT cells whose tuned velocities (v_x, v_y) form an expanding patter. Different MST cells are sensitive to a variety of motion field patterns – not just expanding. I sometimes refer to these as "global" motion patterns because the receptive fields are so big.

The figure below illustrates the receptive field structure of an MST cell. At each location of the receptive field, the cell gets excitory inputs from a (v_x, v_y)-sensitive cell in area MT. Each of the ellipses in the figure illustrates one MT cell. Only about 30 such cells are shown. Each MT cell itself receives excitatory input from a set of V1 cells, namely from those V1 cells whose spatial orientation and normal velocity peak sensitivity is consistent with the velocity of the MT cell. (The MT cell's responses were sketched out last lecture.)

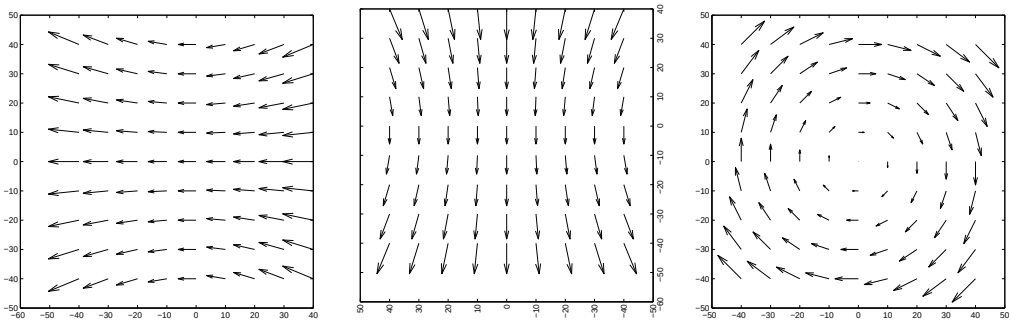


Rotation of viewer

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, or by rotating the eyes within the head, or both a the same time. Note that when the viewer's head rotates, this induces both a translation and a rotation, since the viewer rotates the head around some point in the neck. Lets only concern ourselves with pure rotation for the moment.

When the viewing axes rotates smoothly, a smooth motion field is produced on the retina. If one projects onto an image plane as we did for translation, then one can derive equations for the motion field. These equations depend on the axis that the viewer is rotating around, and on the speed of rotation. These equations are a bit more complicated to derive, so I will skip them and just show pictures.

The two fields on the left show the motion for panning (left and right) and tilting (up and down). The velocity vectors within each of these motion fields are not exactly parallel. The slightly curving of the fields is a subtle artefact that is due to projection onto plane $Z = f$. We can ignore this "second order" detail. The roll field on the right occurs when the rotation axis is the axis in which the viewer are looking. In this case, the speed increases radially away from the axis, like a spinning wheel.



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

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

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

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*. This nerve and other nerves carry information such as blink commands, accommodation controls, and pupil contraction controls. Some of these signals are computed directly in the midbrain and nearby structures, without going to the cortex. This allows very fast feedback to control the eyes. We will discuss an example next.

VOR (vestibulo-ocular reflex)

One fundamental eye movement is called the vestibulo-ocular reflex or VOR. When the head moves – whether it is translation or rotation or both – the motion causes a shift in the retinal image. 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. Look 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 you can move your head quite quickly and maintain your gaze on the object. The VOR plays a central role in this.

The VOR depends on the vestibular system (V) 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*. These are filled with fluid, and when the head rotates, the fluid moves in the canal and this fluid motion is sensed by little mechanical receptors. (Details omitted.) If the head continues to rotate, the fluid drags along and eventually has the same speed as the canal itself. At that point, if the head *stops* rotating, then the fluid keeps going and again the system senses the fluid motion relative to the canal, which sends a (erroneous) signal that the head is rotating again. This is what happens when you spin around 10 times, and then stop spinning. (And you fall down.)

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. If the surface moves upwards, then the stones will press down on the surface and if the surface noves downward, then the stones will press less (like when the elevator goes up or down). In the vestibular system, the "grass" is a set of mechnical receptors and the stones are just that – small stones (called otoliths).

The VOR is extremely fast, and the reason this is possible is that the circuit is so short (see below). 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.

Note 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.

Smooth pursuit

Another important type of eye movement is *smooth pursuit* eye movements. These are voluntary movements that keep a desired object on the fovea. An example is the eye movements that you make when you visually track something moving the world e.g. when you watch a dog walk by. These eye movements are relatively slow. For example, if I move my finger in front of your eye, you can keep your fovea tracking on my finger, but only up to some limited speed.

The reason for the speed limitation is that this smooth pursuit system needs to process the motion. If the image of the object that you want to track starts slipping from your fovea, it means that you are moving your eye too slowly or quickly. Your visual system need to estimate this slippage. This requires that the signal reaches all the way to area MT. That is a few stages of processing just to detect that the eye movement is too slow! The brain also needs to compute the correction and send that signal back to the midbrain where the motor correction can be computed and send to the muscles that control the direction of the eye. (The various 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.)

[ASIDE: The following was only briefly mentioned in the lecture. I include it here to be more complete.]

Note that eye movements (VOR and smooth pursuit) produce rotational components in the motion field. For VOR, the rotational components are meant to cancel out the rotational componets that are due to head motions. If VOR is working properly, then there is no net rotational motion field from head motion + VOR. However, there may still be a rotational component to the motion field from smooth pursuit eye movements. This rotation motion field is added to the translational field, and so if you are walking (translating) while visually tracking some other object (perhaps stationary, perhaps not) then your motion field will be the sum of a translation and rotation field. See the slides for an example.

Disentangling the translation (walking) and rotation (smooth pursuit) component fields would be a difficult computational problem, if the visual system could only rely on visual input to do so. Fortunately, since the visual system controls the smooth pursuit, the system "knows" how the eye is rotating. This information could help to disentangle the translation and rotation components of the field.

