Up to now, we have been considering one type of image information at a time e.g. intensity, stereo, motion, texture, shading. But in real situations, an observer has multiple sources of information available and would like to combine this information. We will use the term 'cue combination' for this. By 'cue', we mean both a particular type of image and scene information, along with a mapping which allows an observer to perceive the scene property given the image property.[1] We have discussed in particular how texture is a cue for slant and tilt; binocular disparity, blur, and motion parallax are cues for depth; shading is a cue for local surface orientation and surface curvature.

Cue combinations

What do we do when we have multiple cues available? Suppose we have two sources of image information which I will simply call I_1 and I_2. These variables represent some image measurement, such as binocular disparity, motion, or a description of a texture. We wish to use I_1 and I_2 to estimate a scene variable S.

Let $p(I_1|S)$ and $p(I_2|S)$ be the likelihood functions for each cue on its own and let $p(I_1, I_2|S)$ be the likelihood function of the two cues together. It is common to *assume* I_1 and I_2 are "conditionally independent":

$$p(I_1,\ I_2 \mid S) \ = \ p(I_1 \mid S)\ p(I_2 \mid S).$$

Intuitively, for a fixed scene, conditional independence says that the value of one image variable I_1 tells us nothing about the value of the other image variable I_2. For example, I_1 might be the sizes of texture elements and I_2 might be the foreshortening of texture elements. Or I_1 might be all the texture cues and I_2 might be the binocular disparities of the texture elements. Note: conditional independence is just a model. In reality, there might be a weak dependence, but we ignore this dependence to keep the model simple.

Suppose the likelihood functions $p(I_1 \mid S = s)$ and $p(I_2 \mid S = s)$ both have a Gaussian shape[2] and with means s_1, s_2 and variances σ_1^2, σ_2^2, respectively.

$$p(I_1 = i_1|S = s) \ = a_1 e^{-\frac{(s-s_1)^2}{2\sigma_1^2}}$$

$$p(I_2 = i_2|S = s) = a_2 e^{-\frac{(s-s_2)^2}{2\sigma_2^2}}.$$

where a_1 and a_2 are constants. If we assume conditional independence, then the likelihood function $p(I_1 = i_1,\ I_2 = i_2 \mid S = s)$ is just the product of these two likelihood functions.

What is the s that maximizes the likelihood $p(I_1 = i_1,\ I_2 = i_2 \mid S = s)$? We next show that the maximum likelihood estimate is a linear combination of the maximum likelihood estimates of the two cues when they are on their own. We want to find the s that maximizes

$$p(I_1|S)\ p(I_2|S) = a_1 a_2 e^{-\frac{(s-s_1)^2}{2\sigma_1^2}} \ e^{-\frac{(s-s_2)^2}{2\sigma_2^2}}$$

[1] Usually the mapping is from scene to image, whereas the vision system wants to map from image to scene, which is why vision is a more difficult problem than graphics!

[2] Recall that this doesn't mean that they are Gaussian probability functions, in the sense that they have unit area. Likelihood functions in general do not integrate to 1 when you integrate over the scene variable $S = s$.

and so we want to minimize

$$\frac{(s-s_1)^2}{2\sigma_1^2} + \frac{(s-s_2)^2}{2\sigma_2^2}.$$

Take the derivative with respect to s and set it to 0. This gives

$$\frac{s-s_1}{\sigma_1^2} + \frac{s-s_2}{\sigma_2^2} = 0$$

and so

$$s = (\frac{s_1}{\sigma_1^2} + \frac{s_2}{\sigma_2^2}) \;/\; (\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2})$$

Note that this is of the form

$$s = w_1 s_1 + w_2 s_2$$

where $0 < w_i < 1$ and $w_1 + w_2 = 1$. In particular,

$$w_1 = \frac{{\sigma_1}^{-2}}{{\sigma_1}^{-2} + {\sigma_2}^{-2}} \qquad w_2 = \frac{{\sigma_2}^{-2}}{{\sigma_1}^{-2} + {\sigma_2}^{-2}}$$

which can be rewritten:

$$w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \qquad w_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

For example, if $\sigma_1 \ll \sigma_2$, then $w_1 \approx 1$ and $w_2 \approx 0$.

This *linear cue combination* method says that if one cue is more reliable than the other, then the more reliable cue should have a heavier weight. The linearity might not be intuitive, however. You might think that a "winner take all" approach would be better, namely that one should put *all* the weight on the more reliable cue. To understand why "winner take all" is wrong, note that we are assuming conditional independence of the cues, which intuitively means that $I_1 = i_1$ and $I_2 = i_2$ give you different information about s. Even though one cue may be more reliable than the other, the less reliable one still gives information and so it should not be entirely ignored.

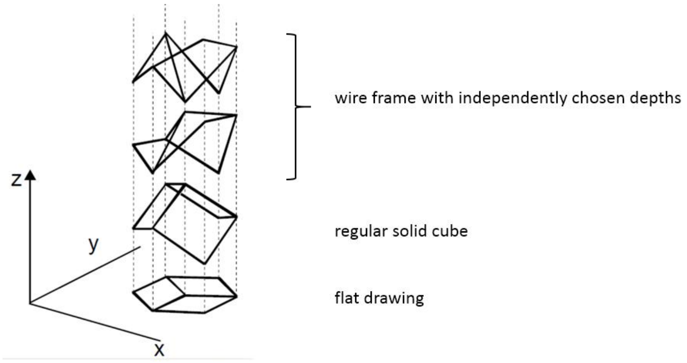
Many experiments have shown that this cue combination theory often does describe human performance. For example, one can vary the noise in one of the cues, and show that the psychophysical thresholds change, as if the vision system were giving less weight to that cue. (Recall that this requires some assumptions, namely that the psychometric curves can be interpreted in terms of likelihood functions.)

Bayes Rule and MAP (Maximum *a Posteriori*) estimation

We have thus far concentrated on the likelihood function $p(I = i \mid S = s)$. One estimates the scene s that produces the image i, such that the probability of the noise that is required to explain the s to i mapping is maximal. One limitation of this maximum likelihood approach is that it ignores the fact that some scenes s have a higher probability of occurring than other scenes.

As an example, consider that an image of a cube can be arise from several different wire frame figures in 3D. One of these figures is a 3D cube, but there are infinitely many others since one can move the depths Z of the points of the cube without changing the image. (Here we are assuming for simplicity that the image formation occurs by a projection that is parallel to the Z axis, but the same projection ambiguity holds if we use perspective projection.)

We can think of the likelihood function as being uniform over s, since any of the scenes s that project to the image shown is as good as any other one in accounting for the image. (We aren't formally writing down a model of the 'noise' here, but we could do so in terms of a slight jittering of each vertex of the wireframe in 3D and its corresponding jittering in 2D.)



Why does the visual system prefer the cube interpretation over any particular other wire frame interpretation? One idea is that cubes have higher probability of occuring in our world than individual complex shapes that happen to look like cubes just because we are viewing them from a particular direction and some accidental alignment. If the visual system takes account of the higher probability of cubes occuring (than random wire shapes that happen to look like cubes), then it would infer a 3D cube when it sees an image that is consistent with a 3D cube.

A more elaborate example is the *Ames Room* illusion. See the videos:
`http://www.youtube.com/watch?v=Ttd0YjXF0no`
`https://www.youtube.com/watch?v=gJhyu6nlGt8`
An Ames room is a 3D room which is viewed in perspective. The room's walls and floor have a 3D trapezoidal shape, but the viewing position within the room is chosen so that the walls and floor have the same image projection as a 3D cube room. We perceive the room as a cube, even though it isn't. And this leads to some strange consequences when there are other objects within the room.

The video shows that two people who are in different places in the Ames room can have quite different perceived 3d sizes. In the first video above, people move in the room and seem to change size as they change position. It is remarkable that the visual system would interpret people as changing size rather than correctly perceive the actual (non-cube) shape of the scene.

Both of the previous examples seem to work because the visual systems prefers a regular shape (cube or room) over a non-regular one. Rather than trying to come with a theory of 'regularity', we will express this idea in terms of probabilities by saying that regular shapes occur more frequently than particular non-regular shapes that happen to look regular. Specifically we can capture this idea by considering the marginal probability $p(S)$ over scenes, and giving a larger value $p(S = s)$ for particular scenes s. This marginal scene probability is called the scene *prior*, and it plays a role in Bayes Rule (or Bayes Theorem) which I will now derive, and which most of you are familiar with since it is commonly taught in basic probability courses.

One can write the joint probability function $p(I, S)$ in terms of conditional and marginal probabilities in two ways:

$$\begin{aligned} p(I, S) &= p(I|S)\ p(S) \\ p(I, S) &= p(S|I)\ p(I). \end{aligned}$$

Equating right sides and isolating $p(S|I)$ gives us *Bayes Rule*:

$$p(S|I) = \frac{p(I|S)\ p(S)}{p(I)}$$

The function $p(S|I)$ is called the *posterior* probability function. It depends on the prior and on the likelihood. The posterior is really what we are interested in: we want to estimate the probability of a scene $S = s$, given an image $I = i$. One often solves for the maximum of the posterior – or *maximum a posteriori*, as its usually called.

Note that the posterior depends on the prior probability $p(I = i)$ of an image i occuring. One typically does not have a model for $p(I)$, and one does not care about $p(I)$. The reason is that one wants to estimate $S = s$, but $p(I = i)$ doesn't depend on any particular $S = s$. One often solves for the maximum of the posterior – or *maximum a posteriori*, as its usually called – one can ignore the dependence on $p(I = i)$. The reason is that one wants to know the probably of a scene $p(S = s)$ *given that* image i has already occurred.

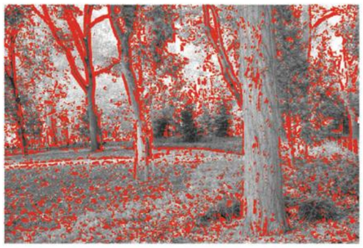
Also note that if the prior $p(S)$ is uniform over S then finding the maximum of the posterior is equivalent to finding the maximum of the likelihood. In many cases, one does not know the prior or one has reason to believe that the prior is relatively flat. In this case one can treat the prior is roughly constant over the region of the parameter S that one is considering.

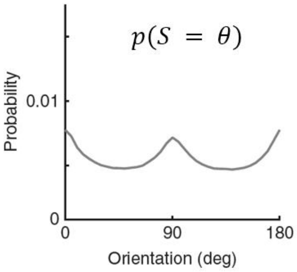
Natural Image and Scene Statistics

Over the past two decades, researchers have begun to collect data and to make quantitative models of image and scene statistics in order to gain insight into the priors and likelihood functions that the visual system seems to use. Image statistics are relatively easy to come by: one takes many images and applies operators such as difference of Gaussian or Gabor filters or others, and fit models to the responses. Scene statistics are more challenging, since they require more sophisticated imaging devies for measuring 3D geometry. But these devices are now available e.g. lidar.

Here I discuss two examples. The first is a study[3] that examined line/edge orientations in natural images. They used computer vision methods to measure the frequency of lines/edges of different orientations. As shown on the right, there are about 50 percent more edges that are vertical or horizontal than are diagonal. This distribution was used to model human percepts of orientation in a psychophysical experiment. It has been known for many years that human observers are better at discriminating orientations that are near vertical or near horizontal than near oblique orientations. This study was able to relate the performance in such orientation tasks to a probability model of likelihoods and priors. (Details omitted.)

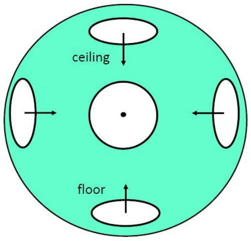
[3] Girshick et al

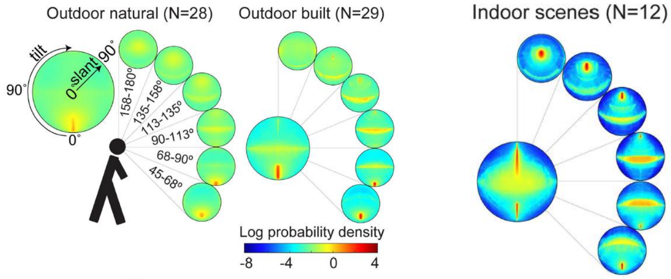




A second example is study[4] of the distributions of slants and tilts of surfaces in various enviroments including outdoor and indoor. Surface depth maps were imaged using lidar techniques (like radar, but uses light). They then fit planes to local surface patches. This gave them the frequencies of different slants and tilts.

One subtlety here is that slant and tilt are measured with respect to some XYZ coordinate system where Z is depth. If one is looking at the ground, then Z will be different than if one is looking upwards towards the ceiling. In the plots below, the slant and tilts are defined with respect to different viewing directionss, specifically elevations. For example, think of the $45 - 68°$ plot as have a Z axis centered at 56 degrees up from the gravity vector and considering the surfaces that are visible in 11 degree neighborhood around the Z axis. They calculate the slant and tilt at each surface point in that neighborhood and they do that for many different scenes. They then used a color map to plot the frequency distribution of slants and tilts. The camera was always at a height close to 2m above the ground so that the statistics correspond to what a typical adult will observe.





Examples of the data are shown above. For example, consider the 90-113 deg elevation which goes from the horizon (parallel to the ground) to 23 degrees above the horizon. For these viewing directions, there is a ridge of peaks for tilts of 0 or 180 deg and all different slants - see the yellow horizontal stripes in the outdoor scenes. This band is presumably due to trees and walls which are vertical surfaces and so the normal is always near perpendicular to the gravity direction. A similar stripe appears at other viewing elevations but the stripe is shifted to other slants and tilts because viewing direction is not horizontal.

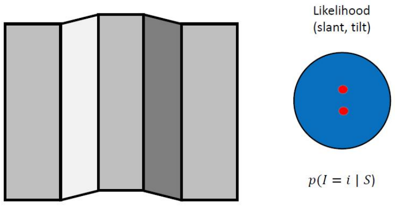
Another example is that in the indoor scenes at elevations above the horizon, there is a hot spot that the corresponds to points on the ceiling (tilt = 90). The slant and tilt of this hot spot shifts with the viewing elevation. There is also a hot spot for floor slants and tilts (tilt = -90 deg) when the viewing direction is in the 45-68 degree range which is below the horizon.

[4] Adams and Elder

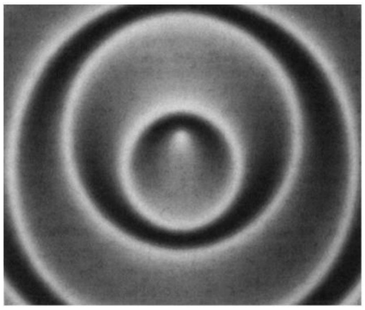
The main point here is that the distribution of slants and tilts in the world is highly non-uniform. The details might turn out to be very interesting for understanding perception, if it turns out that these details correspond to our preferences in perception. It has been shown that the visual system does prefer floor slopes over ceiling slopes, and we will discuss examples below. It hasn't (yet) been shown that there are differences in perception that correspond to the detailed probabilities differences shown the plots above. But perhaps one day that will be shown too.

Depth reversal ambiguity in shape from shading (on a sunny day)

Recall the corrugated plaid illusion from lecture 12. The figure below shows a simplified version of it which can either be interpreted as a ridge (convex) or valley (concave). When we perceive a ridge, the dominant lighting direction is from the left, and the surface is sloped slightly upwards (like a floor). When we perceive a valley, the dominant lighting direction is from the right, and the surface is sloped slightly downwards (like a ceiling). Both of these interpretations are consistent with the image information. We can think of a likelihood function with two corresponding peaks.



Below is a similar example (due to Reichel and Todd 1990). The center region of the shaded pattern can be seen either as a local hill or a local valley. These local curvature percepts depend on seeing the overall surface slant as slightly floor-like or slightly ceiling-like, respectively. Both percepts are valid for the given image. Again we can think of a likelihood function for the surface, which has two maxima corresponding to the two different surface interpretations for this image.



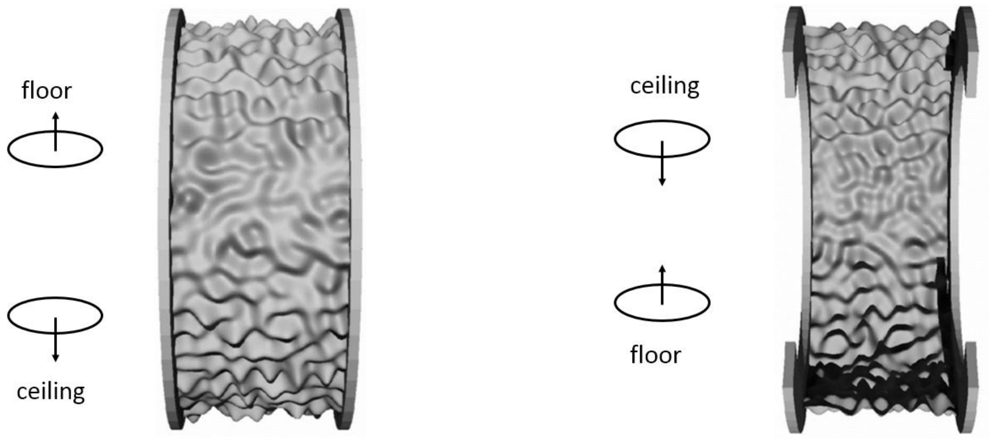
Both of the above are examples of a “depth reversal ambiguity”. See the Exercises for which this ambiguity exists. As we will see next, the visual system often relies on prior assumptions to resolve such two fold ambiguities.

Priors for light from above and global convexity

It has been known for a few hundred years that two-fold ambiguities in shape from shading exist and that the visual system often resolves them by preferring a solution that is consistent with the light being from above. This can be demonstrated informally, as familiar objects such as faces look strange when illuminated from below. It can also be shown formally in shape from shading experiments. Subjects tend to perceive shapes that are consistent with light from above, rather than depth reversed shapes that are consistent with light from below. This prior from light from above is not surprising since, more often than not, scenes are illuminated from above.

Another prior that is well known is surface convexity. We prefer to see individual objects as having a solid shape, that is, overall convex rather than overall concave like a mask. Again this is not surprising since most objects have an overall shape that is solid and hence more convex than concave.

Before I came to McGill as a professor, I did experiments that investigated the prior assumptions $p(S)$ that we use to disambiguate the surface shape in situations of ambiguities. The image classes that I came up with pitted three priors against each other. See below. The surface on the left is overall convex and is illuminated slightly from above the line of sight, and the surface on the right is overall concave and is illuminated from slightly below the line of sight. Surface points that are either just above or just below the center of each image have an overall floor or ceiling slant.



I showed subjects many such images and marked single points on these images, and I asked them say if the points were on a 'hill' or in a 'valley'. Subject's percentage correct scores in each combination of conditions (light direction, floor or ceiling region, global shape) could be modelled as if they were using prior assumptions to disambiguate the two-fold ambiguities. In a nutshell, their percent correct scores were 50 percent plus or minus about 10 percent for each of the three priors. For example, in the floor region for the image on the left, subjects were about 80 percent correct (illuminated from above, floor, overall convex), whereas in the ceiling region in the figure on the right (illuminated from below, ceiling, overall concave) subjects were about 20 percent correct. I've looked at these stimuli thousands of times and I still tend to interpret the hills and valleys using these priors.