Typical scenes contain many objects which are distributed in space, and these objects are often moving and interacting with each other. Because our visual systems sense the light coming from a wide range of directions and analyze the entire image to some extent, we often have the impression that we perceive what is in front of us — all of it. However, if we reflect just a moment on this, we realize that this just isn’t possible, nor it it desirable. Much of what is around us is not of any interest or use. We somehow need to select the information around us that is useful for us. That is the problem of attention. Choosing what to process and what to ignore.

It is common to distinguish ‘Bottom up’ attention processes from ‘top down’ attention processes. Bottom up processes refer to those that are automatic, passive, and stimulus driven. Everything we have discussed in this course so far can be thought of in the way: images are sensed by the eyes and transformed into estimates of depths, orientations, motions, colors, etc. Top down processes are known also to be important, though. These are task- or goal-based, active, and depend on the observer state and knowledge. If we are riding a bicycle, then the task is to visually navigating the space in front of us to safely get to a new position in that space. If we are watching a baseball game, this is still task based. We generally follow the player holding the ball, but sometimes we look at other places to better understand the situations that are developing on the field. I mention this baseball fan example because although it may seem passive to watch sports (relative to playing sports), I want to emphasize that there are important top down aspects too, namely you are trying to figure out what is happening and only processing certain events and locations in your visual field, and you are applying your knowledge of how the sport works. Someone who knows nothing about baseball probably has no idea where they should be looking and doesn’t understand what is going on right in front of their eyes.

The latter idea of ‘top down’ vision is way beyond the scope of this course, and would be better described as visual cognition rather than visual perception. Today when we discuss top down aspects of visual attention, we will be concerned with simple stimuli only. We would like to think about how top down aspects of visual perception are related to concepts that we’ve developed so far about how images are processed in the early stages of the visual pathway. One detail that has come up, and that certainly is relevant to top down perception, is that the pathway from retina to cortex is not entirely one way. It is one way from retina to the LGN, but the path from LGN to the cortex is two way, and indeed there are many more axons from V1 back to LGN than there are from LGN to V1. This feedback loop is almost certainly playing an important role in visual attention. There are also feedback loops in subsequent areas of the visual pathway, from higher areas back to V1.

**Bottom up attention**

There are two main aspects of bottom up attention that I would like to discuss. One is that the visual system computes a set ‘feature maps’ in parallel. These feature maps are retinotopic, just as the LGN and V1 maps are retinotopic. The features that define these maps — luminance, color, contrast, local orientation, size, disparity, motion — are very closely related to the properties of the single cell receptive fields that we have been studying. A key difference in the feature maps we will discuss today is that they are defined psychophysically using human subjects, rather than by electrophysiologically using cats and monkeys. It is assumed that there relationships between the psychophysical feature maps and the electrophysically defined receptive field profiles of single cells. But the goal here is not to nail down these relationships. Rather the goal of attention with
psychophysical experiments using feature maps is to characterize the performance, in particular, to understand what happens when different features are present and how and when they are combined. This relates to the second aspect of bottom up attention that we will examine, namely that parts of the image tend to *pop out*.

Visual pop out experiments commonly use a small number of elements such as colored bars of particular orientation. The paradigm I describe below comes from experiments by Anne Treisman in the 1980s. She measured for example how quickly a red horizontal target bar might pop out in a field of red vertical distractor bars, or how a green vertical *target* bar might pop out in a field of red vertical *distractor* bars.

A more complex case such is shown below. On the left, the distractors are horizontal and vertical red bars. The target that is a green bar. In this case, the target pops out. The idea is that it is a unique item in the color feature map. On the right, the target is a horizontal red bar and distractors are vertical red and green bars. The target pops out because defines a unique item in the vertical bar feature map.

A still more complicated case is shown below. Each display has two different color items (red and green) and two different orientations (vertical and horizontal). Here is where the fun starts. In both of these examples, the unique item (the target) is defined by a unique *conjunction* of features. In particular, *neither feature in is unique in the display*. On the left, half the distractors are red and vertical and the other half are green and horizontal. The target is red and vertical and does not pop out. See if you can find the target on the right.

Treisman found that certain combinations of features pop out and certain ones do not. Moreover, in cases where the target does not pop out, the time taken to say if the item is present or not rises linearly with the number of distractors. It is as if the visual system sequentially searches the display for the odd target. For this reason, Treisman’s experimental paradigm has come to be known as *visual search*. In a typical visual search experiment, hundreds of trials are run and subjects know
what the target is that they are looking for. Despite knowing what you are looking for and lots of practice, the target does not pop out.

![Diagram of visual search examples](image)

Treisman and many others since have explored various aspects of visual search. One aspect is that often there is an *asymmetry* in pop out, for example, the letter C (a curve) pops out in a field of O letters, but an O letter does not pop out in a field of C’s. In the example given in the lecture, a Q like shape pops out in a field of O shapes, but an O shape doesn’t pop out in a field of Q’s. Often such asymmetries occur when one item is *missing a feature* that all of the distractors have (which does not lead to pop out) versus a case in which one item has an extra feature that the distractors do not have (which does lead to pop out). For example, the Q has an extra line relative to O and the C has two line endings relative to O.

A different asymmetry example is that a tilted (diagonal) line pops out from a field of vertical lines, whereas a vertical line does not pop out of a field of tilted lines. One reason this asymmetry might exist is that the visual system has a more accurate representation for vertical lines than for tilted lines. (This has been shown in discrimination experiments, namely our threshold for orientation differences in vertical lines is smaller than for tilted lines.) So the elements of a field of diagonal lines each have larger uncertainty in their orientation, leading to larger overall uncertainty in the field and the vertical line might fall within that greater field of uncertainty. By contrast, vertical lines each have low orientation uncertainty and so the field of vertical lines has an overall low orientation uncertainty, and the diagonal line might fall outside that range of uncertainty. (I am not presenting a model here, but hopefully you can imagine how one could do so.)

Another interesting phenomenon concerns the time that it takes to find the target when the target is present versus when the target is absent. Here we are assuming that the task is to say present versus not present. The main finding is that, for any number of distractors, it takes about half as long to say if the target is present than it takes to say if the target is absent. This should makes sense, if we think of searching for the target. If the target is present and if you search for it using some sort of linear algorithm (scanning the items), then once you find the target you will be able to answer YES. On average, you need to scan about half the display items to find the target when it is present, whereas to determine that the target is NOT present you need to scan all of the items.

This result is more subtle than you might at first think. Subjects in this task do not scan the display items by using a sequence of eye movements. Indeed in some experiments, subjects do not make any eye movements at all. So, what we mean by linearly scanning the targets is unclear. Yet the data consistently show this pattern that the rate at which response time rises with the number of distractors is a factor 2 different when the target is present versus not as if the visual system were linearly scanning for the target.
Computational models: Saliency

The experiments above used very simple, cue-reduced stimuli. These experiments reveal interesting properties such as popout asymmetries and varying reaction time slopes for different features. However, they don’t shed much light on how the visual system processes real images which tend to be more complicated, and whose features are typically more difficult to define.

In the mid-1980’s, researchers (Koch and Ullman, 1985) proposed a model for visual attention, that was much closer to the modern computational models that we see today. In particular, they introduced the term saliency to describe how much an image location is visually prominent or noticeable, i.e. whether attention is drawn to that location. A key property of saliency is that it isn’t tied to a specific feature that pops out. Rather it just means that something at that location causes the visual system to attend there.

The term ‘saliency’ refers to ‘bottom up’ processing. Saliency doesn’t depend on a particular task, or goal target, or single feature map. Rather, saliency refers to a single scalar map – a saliency map – where a big value means very salient and small value means not salient. (Recall that the visual search experiments above, subjects sometimes are told what the target is – namely property or combination of features they were looking for. Saliency models don’t work like that – there is no notion of pre-defined target in the saliency computation.)

Another aspect of the Koch and Ullman model is what happens after saliency is computed. The basic idea is that attention moves to the location of peak salience – called winner-take-all (WTA). The visual system presumably processes the image in greater detail at that location and once that processing is done, it can move on to the next highest salient region, etc.

Itti and Koch (1998) implemented these ideas of the Koch and Ullman model. I will not give all the gory details of Itti and Koch’s model, instead just give the basic steps: compute feature maps for luminance contrast, color, orientation, and disparity and motion if these cues are present. Then run some sort of normalization and lateral inhibition operators to enhance the extrema. This yields a saliency map. The highest saliency location then pops out (‘winner take all’). The system then nulls out the saliency in the neighborhood of this winning location, so that other regions can become salient afterwards (called ‘inhibition of return’). This prevents the visual system from getting stuck on one location.
Most computational models of saliency assume a uniform image resolution – that is, the same acuity and processing for each visual angle. This uniformity assumption is usually made for simplicity, and because the digital images on which the models are tested have uniform spatial resolution.

The human visual system does not have uniform spatial resolution, however, and one might ask why. One reason is that the optics in the visual periphery are not as good as in the center of the image \((x, y) = (0, 0)\), and so the visual system has evolved to have higher sampling density in retinal areas where the image is likely to be in better focus, on average. Interestingly, there is some evidence that the sampling rate in the retina matches the limitations of the optics. That might explain everything. But maybe not.

Another reason why the visual system might not have uniform spatial resolution is that it does not need to. The visual system is used to carry out tasks, which often require information only about particular regions in the scene. Often we can just ignore the rest of the image.

**Eye movements and attention**

We discussed eye movements earlier in the course (lecture 9). Let’s come back to them now in the context of attention. Most models of what drives eye movements are bottom up, driven by salience.
But there are top down factors as well. For example, recall the discussion of the Yarbus experiments where subjects examined the painting of *The Unexpected Visitor* and their eye movement patterns depended very much on what questions they had to answer about the picture.

Eye movements and attention are obviously related but they are not the same thing. One can make an eye movement towards a target and shift attention to that target. But one can also shift attention to a target without making an eye movement. Athletes do this all the time, 'looking' one way but in fact ignoring where the eyes point – in order to 'fake out' an opponent. Drivers also sometimes shift attention without making eye movements. An object in the periphery might attract attention briefly (become salient) for example a face in a crowd, but you might be able to judge without looking at that face whether to make an eye movement or not. For example, you might be able to judge whether the person in the periphery is looking at you (in which case you might want to look at them) or just looking in some direction near you.

Anyhow, the point is that we know from experience that sometimes we shift our attention without making an eye movement, and I argue that this happens often enough that it is something worth studying. One says that an *overt* attention shift is one measured by a movement of the eyes (toward an object to which one attends). A *covert* attention shift is a shift of attention that is not accompanied by an eye movement. How can one study these two types of attention shifts in an experiment? In particular, how can one study covert attention shifts?

**Posner’s cueing paradigm**

Let’s boil this down to a basic vision problem with simple stimuli. The basic question will be: can you detect a briefly presented pattern better if you know in advance where that pattern will appear in the visual field? The idea is that if you know in advance where the briefly presented pattern will appear, then you can covertly shift your attention to that location. Maybe then you would be better at the detecting the pattern i.e. because your attention is at this correct location.

Consider the figure below. There are two types of trials shown: a valid cue trial and an invalid cue trial. Take the valid cue trial first. The subject fixates on a + sign. Then either the left or right square is highlighted briefly - in the figure it is the right square. The square then disappears and shortly after that a stimulus may or may not appear briefly in the interior of the (now disappeared) highlighted square. The subject remains fixated on the + sign, and has to answer the question, such as whether a stimulus indeed appeared on the + sign, and has to answer the question, such as whether a stimulus indeed appeared there or perhaps discriminate two different stimuli that may have appeared there. The highlighted square allowed the subject to shift attention to that location. In particular, Posner found that this cueing leads to a slight improvement in performance, over the case (not shown) in which the square is not highlighted and in which the stimulus could appear either on the left or right. Note that the eyes are not moving in this experiment.

Now consider the invalid cue condition (shown at right in figure below). Here one of the two squares is briefly highlighted, namely the left square. The stimulus then appears at the center of where the right square would be located. The subject has shifted attention to the left which is further away from the stimulus than the original fixation point was. What happens is that the subject’s performance is worse than in the valid case and it is also worse than in a condition (not shown) in which neither square is highlighted. In other words, the invalid cue made performance worse than having no cue at all.
In the above experiment, attention is drawn by a flashing of the square. One might try to understand how attention results from an *exogenous* (external) cue – i.e. the square – by noting that the square causes brain activity in cells whose receptive fields are in its neighborhood, and this activity spreads to nearby cells. In this view, attention is directly linked to this local activity. This is an interesting effect, but perhaps not so surprising.

Posner is credited not for the above experiment, but rather for the following experiment which is more subtle. Rather than using a square to indicate where the stimulus will be, he used an arrow that was located next to the fixation point. This arrow cue is said to be *endogenous* (internal) in the sense that it doesn’t directly create activity where the stimulus will appear. So if subjects were to perform better with the endogenous cue than with no cue, it must be due to a different mechanism than in the exogenous case. Rather it is as if the visual system itself would be priming the cells in that region in a top down fashion. Amazingly, a valid endogenous cue does lead to improvement in performance, and an invalid endogenous cue leads to a decrease in performance.

An example of reaction time data is shown below. Valid trials were used 80 percent of the time, and invalid trials were used 20 percent of the time. (This bias leads the observer to treat the valid cue as indeed ‘valid’.) In a control experiment, the so-called valid and invalid cues conditions were used 50 percent of the time, and since the cues are not reliable here, subjects cannot benefit from them. This control experiment is slightly different than what I described above, since now the square cue does appear but this cue contains no information about where the stimulus will appear. This is a slightly better control than note having a cue at all, since it is closer to the actual experiment.

Many researchers have used Posner’s cueing paradigm to study the relation between covert attention shifts (without or before eye movements) and overt attention shifts (that occur after eye movements). One interesting and surprising result is that covert attention shifts are much faster.
than overt attention shift, namely covert leads overt by 100 ms. When you make an eye movement, attention shifts to the target before the eyes do!

Another interesting result is that – in some technical sense which I omit here – we can attend equally well in the periphery as we can in fovea. This goes against the intuitive assumption that we can attend better to the things we are looking at (i.e. foveating on). Even if we do attend more to things we are looking at, we are still capable of attending equally well in the periphery.

Yet another result worth mentioning is that when we shift our attention from fovea to periphery - regardless of whether we will make an eye movement – an attention window travels over the region in between. So, for example, you can show that you are better at detecting objects at certain times and at certain locations between A and B when attention shifts from A to B. It is as if there is spotlight of attention that moves from A to B, and the visual system attends to regions that are 'illuminated' by the spotlight, when the spotlight passes over them.

One final result concerns the spatial location(s) that we attend to. Consider the example below. Suppose that the target appears at A or B the vast majority of the times. With an endogenous cue of the red arrow, subjects can perform whatever task better when the target appears at the valid location A than at the invalid location B. But suppose very occasionally a target were placed at C rather than A or B. You might think that because distance(A,C) = distance(B,C), subjects should have the same performance at C as at B. However, what happens is quite different in that performance at C is quite similar to A. What seems to be happening is that the long rectangles group A and C positions together as if these regions form an object and as if subjects are attending to the object rather than simply attending to the location A.

Change blindness, inattention blindness

Please see the self-explanatory video links in the slides. (For change blindness, it often takes 10-30 seconds to find the difference between the two images.) These examples reinforce the fact that often we do not perceive what is 'in front of our eyes'!