Perceptual Organization

An image is an array of intensities (RGB), but we don’t perceive the individual points. Instead we perceive groups, parts, hierarchies, etc. How does the visual system organize an image into coherent parts? There are many ways to answer this question, and it depends on what level you choose. Are you describing neurons? Are you describing behavior in a psychophysics experiment? Are you describing an abstract computational model?

Nearly 100 years ago, a group of psychologists in Germany tried to formulate a set of rules for how the visual system organizes the parts of an image. These have become known as Gestalt grouping rules. (The German word ‘gestalt’ means shape or form.) I’ll list several of the rules here. See the slides or illustrations.

- **Prominence**: items that are close to each other in the image should be grouped more than items that are further

- **Good continuation**: items that fall along smooth curves should be grouped together more than items that don’t; think of following a curve up to a point and then predicting where the next part of the curve will be; more likely it will continue smoothly e.g. a straight line will continue to be straight; an arc of circle will continue to be a bigger arc of circle

- **Similarity (color, luminance, size, orientation)**: items that are similar are more likely to be grouped

- **Closure**: incomplete objects tend to be perceived as complete by closing gaps (contour) or filling missing parts of a shape (solid). An interesting example is the image below, called the Kanizsa triangle.

![Kanizsa Triangle Image](image-url)

There are six different shapes in the image, namely three ‘pac-men’ and three pairs of lines. We tend to perceive a background occluded triangle, and a illusory foreground occluding triangle which partly covers both the background triangle and the pac men (five shapes). We perceive these triangles in particular despite their contours being incomplete. Parts of the background triangle are occluded, and the only parts of the foreground triangle that are visible are the edges defined by the pacmen. Amazingly, the remaining ‘edges’ of the foreground triangle seem to be well defined in the image even though there is no change in intensity there.
Convexity: We tend to group items into convex figures rather than concave. For example, on the left below, there is a competition between two smaller partly occluded convex figures versus one larger bowtie-like figure that is not convex. On the right we can perceive one partly occluded convex figure.

A version of such stimuli was used by Zili Liu in an interesting depth from binocular disparity experiment. He placed the upper and lower parts of such figures at different bincular disparities and asked subjects to discriminate the depths. Subjects were much better at the task for stimuli on the left. Liu reasoned that they didn’t group the two parts into one object (at one unique perceived depth).

Classical gestalt grouping laws assume that the things to be grouped are given. But where do the things to be grouped come from? Palmer and Rock in the 1980’s proposed a principle that they called 'uniform connectedness'. The idea is that any connected region that is defined by a uniform property will be perceived as an image region. For example, in the Kaniza triangle figure above when we said that there were six original shapes in the image (the three contour pairs and the three pacmen), we already were assuming these shapes had been grouped – as they are connected regions of black 'pixels').

The 'uniform' constraint of uniform connectedness isn’t just color. It can be any uniform property of the image across some connected region. So, for example, dots of some uniform density or line elements of some uniform orientation both can define a region.

Psychophysics of contour grouping

When the visual system groups a region into an object, it defines points as being either in the interior or exterior of the object, and it defines an object boundary. The grouping rules above define rules both for region grouping and contour grouping. For the moment, let’s consider just contour grouping.

An experiment by Field (1993) used stimuli such as below on the left, and tried to quantify the good continuation and proximity cues. The image shows a set of cosine Gabors of random...
orientation and position. Within this set is a contour defined by a sequence of Gabors. Does it pop out for you? Hint: it is on the left side.

The figure on the right indicates Gabors that would or wouldn’t group well (pairwise, that is). Take the middle Gabor which is horizontal. The three Gabors on the right each would group well with the middle one because a simple smooth contour (roughly a circle) passes through them, such that the tangent to the contour is in the direction defined by the Gabor. The three on the left would not group as well with the middle one because there are no smooth contours that join the horizontal Gabor with any of three on the left, such that the tangent of the contour aligns with the Gabor orientations.

Field’s experiments were in the early 1990s. More recently there has been work trying to relate contour grouping cues to statistics of natural images. Consider two edge fragments defined by position and orientation \((x, y, \theta)\) and define angles \(\theta_{ij}\) as shown. Elder considered a set of natural images and had subject trace contours in the image. They then looked a various statistics on the parameters shown (and other parameters including the luminances on the two sides of the edges – not shown). They found strong correlations between the \(\theta_{ij}\) angles, which is consistent with the idea of good continuation as perfect correlation would mean that the edge fragments always lay on circles.

Elder developed a mathematical model of the posterior probability that two edges belong to the same contour. I’ll omit details of the model here, except to say that it didn’t just consider the good continuation cue; it also considered proximity and other cues. An example of results are shown above right. There is a horizontal bar at \((x, y) = (0, 0)\) and at other locations all orientations are considered and their posterior probability is indicated by a colored bar. The probability needed to be above some threshold to be shown, which is why some are bowties.

The computational model assumes directed oriented edges, starting from the edge at \((0, 0)\) and going to the right. So an oriented edge on the on the left side would have to group with the horizontal contour at \((0, 0)\) via a long circular arc. This is why the figure is not symmetric i.e. the red (hot) spots are on the right only.

Another observation is that the hot spots bowties don’t see to be centered on co-circular arcs. Rather, they seem slightly rotated towards the horizontal. This suggests that there is a competing tendency for edges to be parallel (both horizontal).
Perceptual organization and 'mid-level' vision

Thus far we have been concerned mostly with 2D grouping. Given an image, the visual system forms contours and regions using principles just discussed. Some of the examples involved occlusions, but we discussed them only briefly. Let's next briefly consider how to interpret these 2D contours and regions in terms of 3D scene properties e.g. depth, material, illumination, transparency.

The figure below left (from Adelson) shows a shaded rectanguloid image which is decomposed into a reflectance image and an illuminance image. The reflectance image shows what the surface would look like if the illumination were uniform, and the illumination image shows what the surface would look like if the reflectance were uniform. The bright top is due to illumination from above. The idea here is that when our visual systems perceive the 3D rendering in the upper image, we are explaining the intensities in terms of reflectance and illumination.

This decomposition is an example of 'mid-level' vision. We say 'mid-level' because is requires that 'low level' 2D regions and contours have already been extracted, and it comes before 'high-level' vision which is often associated with object recognition ("its a bird, its a plane, no its ... ").
Mid-level vision remains one of the most poorly understood yet most fascinating aspects of visual perception. It is believed the *edge junctions* play an important role in mid-level vision. In the figure on the above right, various junctions \((X, L, T, Y, \Psi)\) are indicated. Different junctions have more or less likely interpretations in terms of the 3D events—such as occlusions—that may have given rise to them.

For example, take the shadow shown below on the left which I have abstracted out on the right. The shadow edge cuts across a reflectance edge (pavement to grass). The image intensity is defined by the product of illumination and reflectance, \(I(x, y) = E(x, y)R(x, y)\) and both the latter two have an edge discontinuity. Such a situation clearly defines an X junction.\(^1\)

Note that another interpretation of the X-junction is a transparency layer (like a sunglasses lens) with an edge cutting across a background scene edge. An example is shown below on the right. On the left are examples of T junctions which often indicate occlusions. The top of the T belongs to the occluder, and the stem of the T belongs to the occluded surface.

---

\(^1\) One can show that certain only certain combinations of intensities in the four regions are consistent with a shadow and reflectance edge. I omitted the details this year.
T-junctions play a role in White’s illusion. The short white bars on the left are either painted onto the black bars or they belong to a transparent foreground rectangle. The short dark grey bars on the right belong to a background rectangle that is occluded by the long black bars. (Keep in mind that this is an illusion! The short bars on the left and right are the same grey level.) Why exactly the grey bars on the left appear lighter than the ones on the right is not entirely known, and I will not attempt to explain it here. The main point here is that the X and T junctions are interpreted in terms of reflectances, and occlusion relationships between the regions. If you want to see more examples like the above, see http://persci.mit.edu/gallery/lightness_illusions.

One more example: one of my favorites. The image below appears to show four black disks and four white disks. Surprisingly, the intensities within each of the disks on the left are exactly the same as the intensities within the disks on the right.

This is only a surprise if you believe that we perceive image intensity. We don’t. Rather – at least, when it can – our visual systems constructs an explanation of the intensities in terms of what is happening in the scene. When the intensities depend heavily on a combination of illumination and reflectance, or in the case of the disks – reflectance and transparency, what we perceive is the properties of the surfaces (or lights, or fog) of the scene. The scene on the left has a bright background and the fog attenuates and scatters this background light. The scene on the right as a black background, and the fog scatters and attenuates the light from the radiating white disks.

So, is this an illusion? It depends. On the one hand, I would argue this is no more an of an illusion than what we experience every day, in a scene that has both shadows and directly illuminated areas. We do not perceive the luminance values of light reflected from surfaces, but rather we perceive some combined brightness value that depends on the material of the surface, the illumination, and these same parameters of the surrounding surfaces. On the other hand, it is an
illusion in the sense that when you point out that the disks are the same on left and right, people are surprised and delighted, as if the image has tricked them somehow – just like with other visual illusions.

From perceptual organization to object recognition

The image below shows a Dalmation dog, sniffing in the grass. Usually the first time one sees this image, it takes some time for the dog to appear. However, once that happens, the dog appears immediately – even if you haven’t seen the image for years.

To be honest, we have no idea how the visual system can pull out a dog from an image like that. Even if the objects are well isolated such as a silhouette image, object recognition is a difficult problem. Why is object recognition so difficult? There are many reasons: (1) an object can be at any position, size, pose; (2) the illumination can vary, so the intensity patterns on the object will vary too; (3) the object can be partly occluded or it can be camouflaged, such that it has the same luminance or texture as the background.

Another reason object recognition is difficult is that its unclear what we mean by 'object', since objects can be classified or recognized at many levels. Is something an animal, or a dog, or a Beagle, or it is my Beagle? The same 'level' issue arises for any other object: is something a vehicle, car, Volvo XC 70, or my car? Theories of levels of recognition have been developed but we only have so much time left today, and I would prefer to tell you about computational theories of recognition.

Theories of object recognition: from Marr to machine learning

When I began studying vision, I came across a well known book by David Marr. Marr wrote a number of extremely influential papers in the 1970’s, one of which was on object recognition. The theory is no longer taken seriously as a candidate for how humans visual recognition works, but at
the time it made sense. So I’ll briefly describe the theory and why it made sense. Then I’ll tell you why most people now think it doesn’t make sense.

Let’s take an example. When you think of an object such as horse, you imagine an animal with a barrel shaped body, a thick hind, muscular shoulders, and long neck and head. A horse is quite different from a dog and from a giraffe, and the main differences is the shape. To recognize an image of a horse, it seems natural to transform that 2D image into the 3D shape that generated it, and then match the representation of the 3D shape to a representation that is stored in memory. This representation would include the various parts of the horse and how they fit together.

The pipeline below shows the basic idea. The visual system first extracts lines, edges, textures, etc from the image. It then takes these 2D image properties and uses Shape from X methods to infer surfaces properties: relative depth of points, slant and tilt, and bounding contours. From these surface representations, the visual system constructs a 3D model that fits the surface. This 3D model might be built out of primitives such as cylinders, spheres, and more general versions of these shapes e.g. that allow for bending and tapering. Then this 3D model is matched to 3D models stored in memory. The pipeline below shows the basic idea.

The models themselves might be hierarchical e.g. an arm consists of an upper arm, forearm, and hand. A hand consists of the palm, fingers, etc. Moreover, different types of objects might be organized into hierarchies e.g. a dog, horse, and giraffe are all four legged animals and they differ visually mainly in the shape of their parts. Marr discussed these issues of hierarchies, but never developed\(^2\) as part of his computational theory. He spent more of his effect on the early stages of the pipeline.

There are several significant problems with Marr’s theory and by the early 1990’s people started moving away from it. The main problem is that often there is not enough information from Shape from X cues to accurately estimate the slant, tilt, and bounding contour, and in turn there is not enough information to form a 3D model. Yet we can recognize objects in these situations. People felt that something else must be going on.

Supporters of Marr’s theory argued that our visual systems do often extract 3D information. Even though something else must be going on sometimes (recall the Dalmation example), many times in real 3D viewing situations where we have all the cues (stereo, motion, detailed texture, etc), our brains do reconstruct 3D scenes. How else could you sink a basket through a hoop from 10 meters away? How else could you run along a rocky path and know where to place your feet?

\(^2\)Marr died at age 35 in the early 1980’s.
How else could you reach into the back of the fridge for the ketchup without your hand knocking your hand against the other objects in the way?

The counter argument is that, yes we can judge depth and 3D shape sometimes e.g. to grasp an object, but this does not imply that 3D shape representations are used for object recognition. There is alterative way to do recognition, which has proven to be much more successful over the past 20+ years. This method is crudely illustrated below.

The visual system analyzes the image and forms representations of the local intensity structure. These could be raw image intensities, or they could be patches of outputs from a convolution. For example, the visual system might represent the intensity gradients in a local patch and use this gradient map or even collapse the gradients into a histogram (ignoring the spatial positions of the gradients with the patch). These methods and ‘theories’ have become performance driven, and make heavy use of machine learning techniques. This approach has become even more feasible in the past decade, now that any researcher has access to millions of images on the web, and large scale computation is relatively cheap.

Object recognition using machine learning methods is a topic the merits its own course, and I will not attempt to give any further details. Besides, this is a course in computational models of human vision, and most of the methods I just alluded to are pure computer vision methods. Some researchers believe that these models are closely related to how humans must solve the recognition problem, but the link between the methods and the human psychophysical data isn’t there yet. At the end of the lecture, I discussed a recent paper by Ullman that examined human performance at an object recognition task, and compared performance to that of various computational models that have been proposed – some pure computer vision, and some that allegedly model how the brain solves the recognition problem. The specific aspect Ullman examine is how much one could crop images before people and algorithms failed to recognition them. In a nutshell, Ullman found that the performance of people was quite dissimilar from that of all of these algorithms, suggesting that these algorithms are missing something very important about humans solve the problem. (See the slides for example images.)