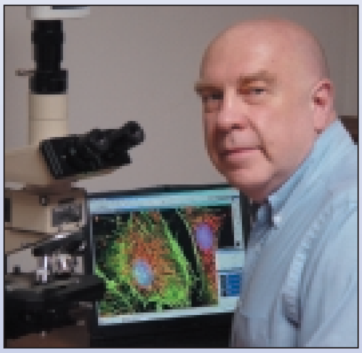
Article



John Russ is the author of The Image Processing Handbook, Computer Assisted Microscopy, Practical Stereology, Forensic Uses of Digital Imaging, Image Analysis of Food Structure, as well as many other books and papers. He has been involved in the use of a wide variety of microscopy techniques and the computerized analysis of microstructural images for nearly 50 years. One of the original founders of Edax International (manufacturer of X-ray analytical systems), and the past Research Director of Rank Taylor Hobson (manufacturer of precision metrology instruments), he has been since 1979 a professor in the Materials Science department at North Carolina State University. Now retired, he continues to write and lecture on topics related to image analysis.

Seeing the Scientific Image

John C. Russ

Materials Science and Engineering Department,
North Carolina State University, Raleigh, NC

What we see and why

Human beings are intensely visual creatures. Most of the information we acquire comes through our eyes (and the related circuitry in our brains), rather than through touch, smell, hearing or taste. For better or worse, that is also the way scientists acquire information from their experiments. But the skills in interpreting images developed by millions of years of evolution don't deal as well with scientific images as they do with "real world" experiences. Understanding the differences in the types of information to be extracted, and the biases introduced by our vision systems, is a necessary requirement for the scientist who would trust his or her results. It is the purpose of this article to acquaint or remind readers of the needs and remedies.

The percentage of information that flows through visual pathways has been estimated at 90-95% for a typical human without any sensory impairment. Indeed, our dependence on vision can be judged from the availability of corrective means - ranging from eyeglasses to laser eye surgery - for those whose vision isn't perfect or deteriorates with age. Hearing aids and cochlear implants are available (but under-utilized) for those with severe hearing loss, but there are no palliatives for the other senses. As taste becomes less sensitive, the only solution is to sprinkle on more chili powder.

Not all animals, even all mammals, depend on or use sight to the extent that we do (Figure 1). Bats and dolphins use echolocation or sonar to probe the world about them. Pit vipers sense infrared radiation. Moles, living underground, trade sight for sensitive touch organs around their nose. Bloodhounds follow scents and butterflies have taste organs so sensitive they can detect single molecules. Some eels generate and sense electric fields that interact with their surroundings. Fish and alligators have pressure sensors that detect very slight motions in their watery environment. Birds and bees both have the ability to detect the polarization of light, as an aid to locating the sun position on a cloudy day. Birds and some bacteria seem to be able to sense the orientation of the earth's magnetic field, another aid to navigation. And many birds and insects have vision systems that detect infrared or ultraviolet colors beyond our range of vision.

It is not easy for humans to imagine what the world looks like to a bat, eel or mole. Indeed, even the word "imagine" demonstrates the problem. The root word "image" implies a picture, or scene, constructed inside the mind. Dependent as we are on images, that is the only organization of world data that is comprehensible to most of us, and our language reflects (sic!) or illustrates (sic!) that bias.

With two forward facing eyes capable of detecting light over a wavelength range of about 400-700 nm (blue to red), we are descended from arboreal primates who depended on vision and stereoscopy for navigation and hunting. Many animals and insects instead sacrifice stereoscopy for coverage, with eyes spaced wide to detect motion. A few, like the chameleon, can move their eyes independently to track different objects. But even in the category of hunters with stereo vision there are many birds with much better sight than humans. Eagles have resolution that can distinguish a mouse at a range of nearly a mile. In fact, most birds devote a much larger portion of their head space to eyes than we do. In some birds, the eyes are so large that it affects other functions, such as using blinking to force the eyes down onto the throat to swallow food.

An oft-quoted proverb states that "a picture is worth a thousand words," and is used as an illustration of the importance of images and their apparent rich information content. But the proverb is wrong in many ways. First because a typical image, digitized and stored in a computer, occupies the space of several million words of text (and even then, the resolution of modern digital cameras is far less than that of the human eye, which has about 160 million rods and

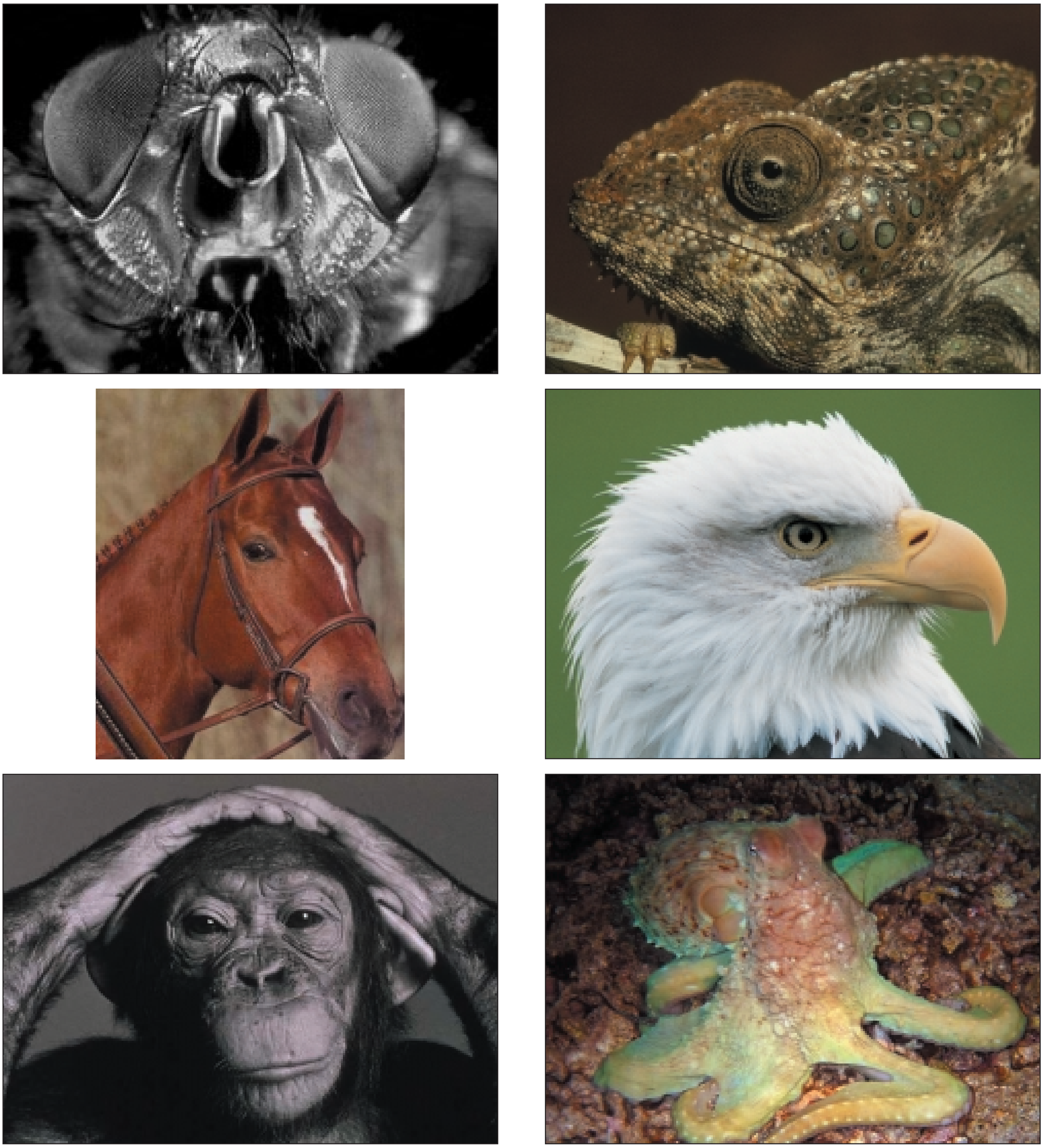


Figure 1. Eyes come in many forms, optimized for different purposes. Insect eyes consist of many individual lenses and sensors, producing comparatively low resolution. The chameleon can swivel its eyes independently to track different objects in left and right visual fields. The horse has little stereo vision but a broad field of view. The eagle's acuity and resolution is extremely high. Primates are well adapted for stereo vision and also have greater sensitivity to red colors than most other animals. The eye of the octopus apparently evolved independently and has its neural circuitry on the opposite side of the retina, but provides very good acuity and color sensitivity.

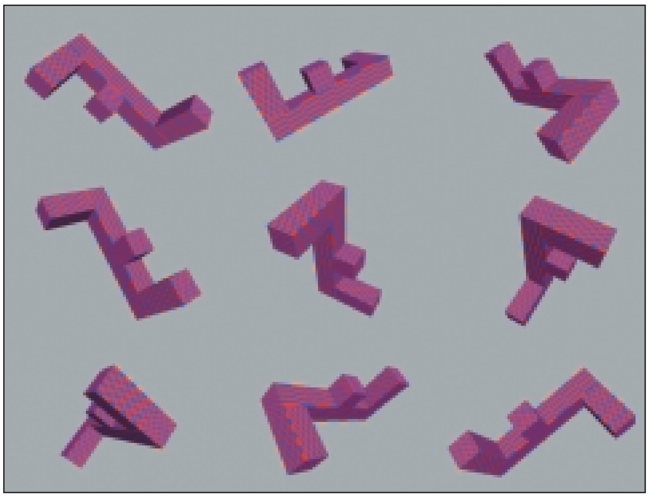


Figure 2. Some of these objects are identical and some are mirror images. The length of time required to turn each one over in the mind for comparison is proportional to the angular difference.

cones). Second, because as a means of communicating information from one person to another the image is very inefficient. There is little reason to expect another person to derive the same information from a picture as we did without some supporting information to bring it to their attention and create a context for interpreting it. Arlo Guthrie describes this in "Alice's Restaurant" as "Twenty-seven 8×10 color glossy pictures with circles and arrows and a paragraph on the back of each one." And that is not a bad description of many typical scientific papers!

Human vision can extract several different kinds of information from images, and much of the processing that takes place has been optimized by evolution and experience to perform very efficiently. But at the same time, other types of information are either ignored or suppressed and are not normally observed. Sherlock Holmes often criticized Watson for "seeing but not observing" which is as good a distinction as any between having photons fall upon the retina and the conscious mind becoming aware. We will examine some of the processes by which information is extracted and the conscious levels of the mind alerted, and note that the extraction process overlooks some kinds of information or makes them very difficult to detect.

Recognition

The goal of much of human vision is recognition. Whether searching for food, avoiding predators, or welcoming a mate, the first thing that catches our attention in an image is something familiar. To be recognized, an object or feature must have a name - some label that our consciousness can assign. Behind that label is a mental model of the object, which may be expressed either in words, images or other forms. This model captures the important (to us) characteristics of the object. It is unfortunate in many scientific experiments that the task assigned to human vision is not the recognition of familiar objects but the detection and description of unfamiliar ones, which is far more difficult.

The basic technique that lies at the root of human vision is comparison. Nothing in images is measured by the eye and mind; we have no rulers and protractors in our heads. Features that can be viewed next to each other with similar orientation, surroundings and lighting can be compared most easily. Ones that must be mentally flipped or rotated are more difficult. Figure 2 shows an example in which the length of time required to mentally turn each object over in the mind to match alignments and determine which features are the same, and which are mirror images, is proportional to the angular differences between them. Comparisons to memory work the same way, and take time. If the remembered object is a very familiar one, then the underlying model consists of a set of characteristics that can be compared. That is, after all, how recognition works.

If the remembered object is not familiar, and has no label and model, then comparison depends on just which characteristics of the original view were remembered. How well the memory process worked, and which features and characteristics were selected for recall, are themselves subject to comparisons to still other models. As an example, eyewitness accounts of crime and accident scenes are notoriously unreliable. Different observers select different attributes of the scene or suspect as being notable based on their similarity or difference from other objects in memory, so of course each person's results vary. Police sketches of suspects rarely match well with actual photographs taken after capture. In some respects they are caricatures, emphasizing some aspect (often trivial) that seemed familiar or unusual to an individual observer.

A threshold logic unit implements the process that can signal recognition based on the weighted sum of many inputs. This process may not duplicate the exact functions of a real neuron, but is based on the McCullough and Pitts "perceptron" model which successfully describes the overall process (Figure 3).

Recognition is frequently described in terms of a "grandmother cell." This is a theoretical construct, not a single physical cell someplace in the brain, but it provides a useful framework to describe some of the significant features of the recognition process. The idea of the grandmother cell is that it patiently examines every image for the appearance of grandmother, and then signals the conscious mind that she is present. Processing of the raw image that reaches the retina proceeds in several places, including the retina and visual cortex, and in a very parallel fashion. In the process, several characteristics that may roughly

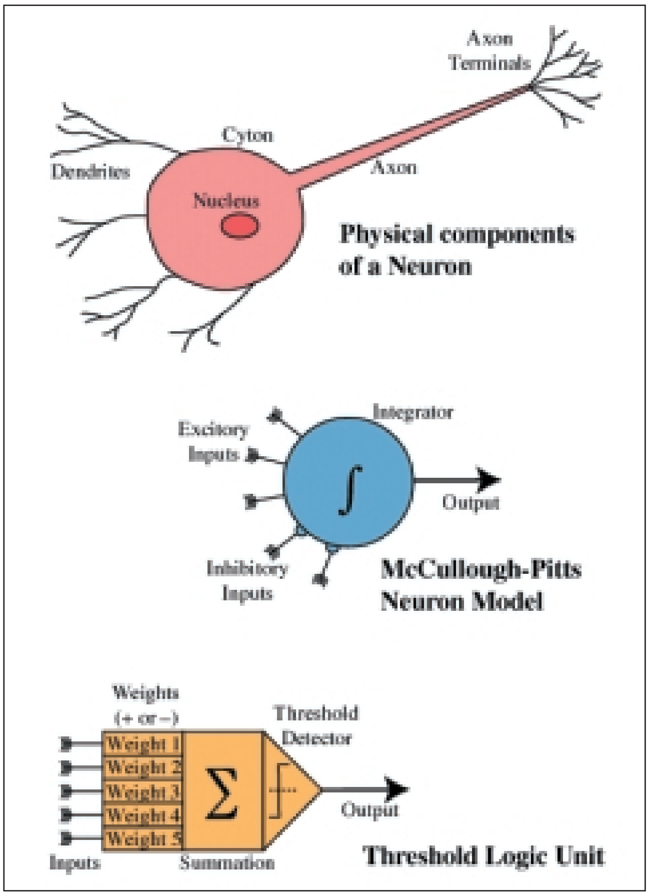


Figure 3. Comparison of a physical neuron, the McCullough and Pitts simplified model of a neuron, and its implementation as a threshold logic unit. If the weighted sum of many inputs exceeds a threshold then the output (which may go to another logic unit) is turned on. Learning consists of adjusting the weights, which may be either positive or negative.

be described as color, size, position and shape are extracted. Some of these can be matched with those in the stored model for grandmother (such as short stature, white hair, a smile, perhaps even a familiar dress). Clearly, some characteristics are more important than others, so there must be weighting of the inputs. If enough positive matches exist, and in the absence of negative characteristics (such as a flaming red mustache), then the "grandmother" signal is sent.

This simple model for a "threshold logic unit" evolved into the modern neural net, in which several layers of these individual decision making units are connected. Their inputs combine data from various sensors and the output from other units, and the final output is a decision, basically recognition that the inputs match some recognized circumstance or object. The characteristics of neural net decisions, whether performed on images in the human mind or other types of data in a computer circuit, are very high speed (due to the extremely parallel way that all of the small logic decisions happen at the same time), the ability to learn (by adjusting the weights given to the various inputs), and the tendency to make mistakes.

Everyone has had the experience of thinking they recognized someone ("grandmother") and then on closer inspection realized that it isn't actually the right person at all. There were enough positive clues, and the absence of negative ones, to trigger the recognition process. Perhaps in a different situation, or with a different point of view, we wouldn't have made that mistake. But setting the threshold value on the weighted sum of positive inputs too high, while it would reduce false positives, would be inefficient, requiring too much time to collect more data. The penalty for making a wrong identification is a little minor embarrassment. The benefit of the fast and efficient procedure is the ability to perform recognitions based on incomplete data.

In some implementations of this logic, it is possible to assign a probability or a degree of confidence to an identification, but the utility of this value depends in high degree upon the quality of the underlying model. This may be represented as the weights in a neural net, or the rules in an fuzzy logic system, or in some other form. In human recognition, the list of factors in the model is not so explicit. Writing down all of the characteristics that help to identify grandmother (and the negative exclusions) is very difficult. In most scientific experiments, we try to enumerate the important factors, but there is always a background level of underlying assumptions that may or may not be shared by those who read the results.

It is common in scientific papers that involve imaging to present a picture, usually with the caption "typical appearance" or "representative view." Editors of technical journals understand that these pictures are intended to show a few of the factors in the model list that the author considers particularly significant. But of course no one picture can be truly "typical." For one thing, most naturally occurring structures have some variability and the chances of finding the mean value of all characteristics in one individual is small, and in any case would not demonstrate the range of variation.

But even in the case of an individual like grandmother, no one picture will suffice. Perhaps you have a photo that shows her face, white hair, rimless glasses, and she is even wearing a favorite apron. So you present that as the typical picture, but the viewer notices instead that she is wearing an arm sling, because on the day you took the picture she happened to have strained her shoulder. To you, that is a trivial detail, not the important thing in the picture. But the viewer can't know that, so the wrong information is transmitted by the illustration.

Editors know that the real meaning of "typical picture" is "this is the prettiest image we have." Picture selection often includes an aesthetic judgment that biases many uses of images.

Technical specs

The human eye is a pretty good optical device (Figure 4). Based on the size of the lens aperture (5×10^{-3} m) and the wavelength of light (about 5×10^{-7} m for green) the theoretical resolution should be about 10^{-4} radians or 1/3 arc minute. The lens focuses light onto the retina, and it is only in the fovea, the tiny portion of the retina in which the cones are most densely packed, that the highest resolution is retained in the sensed image. One arc minute is a reasonable estimate for the overall performance of the eye, a handy number that can be used to estimate distances. Estimate the size of the smallest objects you can resolve, multiply by 3000 and that's how far away you are in the same units. For example a car (about 13 feet long) can be seen from an airplane at 40,000 feet, and so on.

The number of 160 million rods and cones in the retina does not estimate the actual resolution of images. When we "look at" something, we rotate our head and/or our eyeballs in their sockets so that the image of that point falls onto the fovea, where the cone density is highest. The periphery of our vision has relatively fewer cones (which respond to color) as compared to rods (which sense only brightness), and is important primarily for sensing motion and for judging scene illumination so we can correct for color balance and shading. To produce a digital camera that captured entire scenes (as large as the area we see) with resolution that would match human foveal vision, so we could later look anywhere in the stored image and see the finest detail, would require several billion sensors.

Human vision achieves something quite miraculous by rapidly shifting the eye to look at many different locations in a scene and, without any conscious effort, combining those bits and pieces into a single perceived image. There is a blind spot in the retina without sensors, where the optic nerve connects. We don't notice that blind spot because the brain fills it in with pieces interpolated from the surroundings or stored from previous glances. Tests in which objects appear or disappear from the blind spot prove that we don't actually get any information from there - our minds make something up for us.

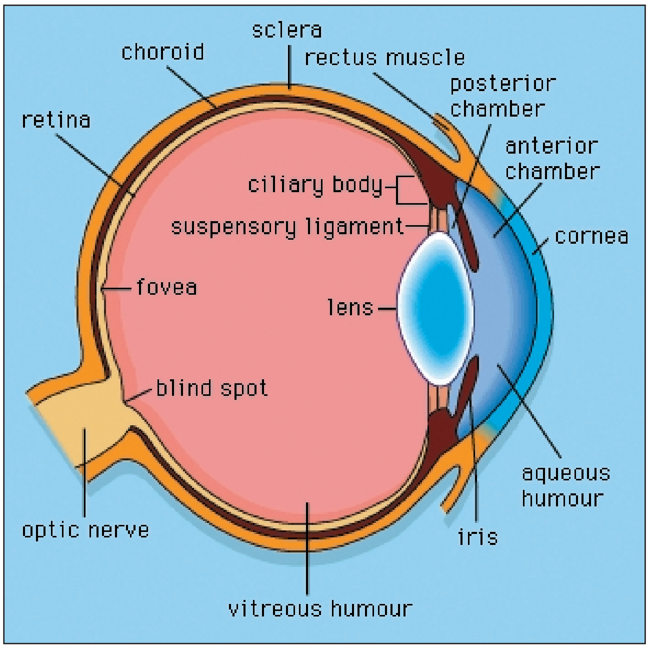


Figure 4. Simplified diagram of the eye, showing the lens, retina, fovea, optic nerve, etc.

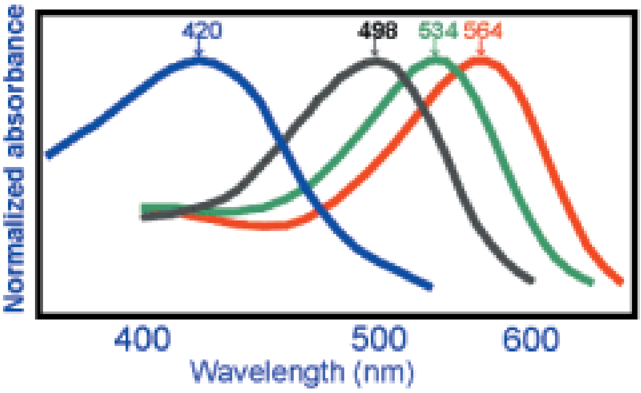


Figure 5. Sensitivity of the rods (shown in grey) and three kinds of cones (shown in red, green and blue) as a function of wavelength. Human vision detects roughly the range from about 400 nm (blue) to 700 nm (red).

The eye can capture images over a very wide range of illumination levels, covering about 9 or 10 orders of magnitude from nearly single photon performance on a starlit night to a bright sunny day on the ski slopes. Some of that adaptation comes from changing the aperture with the iris, but most of it depends on processing in the retina. Adaptation to changing levels of illumination takes some time, up to several minutes depending on the amount of change. In the darkest few orders of magnitude we lose color sensitivity and use only the rods. Since the fovea is rich in cones but has few rods, looking just "next to" what we want to see ("averted vision") is a good strategy in the dark. It shifts the image over so to an area with more rods to capture the dim image, albeit with less resolution.

Rods are not very sensitive to light at the red end of the visible spectrum, which is why red light illumination is used by astronomers, submariners, and others who wish to be able to turn off the red light and immediately have full sensitivity in the dark-adapted rod vision. The cones come in three kinds, which each respond over slightly different wavelength ranges (Figure 5). They are typically called long, medium and short wavelength receptors, or, more succinctly, red, green and blue-sensitive. By comparing the response of each type of cone, the eye characterizes color. Yellow is a combination of red and green, magenta is the relative absence of green, and so on.

Because of the different densities of red, green and blue sensitive cones, the overall sensitivity of the eye is

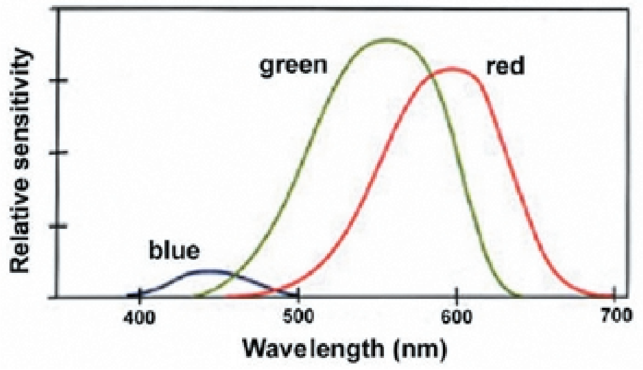


Figure 6. Overall visual sensitivity is lowest for blue light, highest for green.

greatest for green light and poorest for blue light (Figure 6). But this sensitivity comes at a price: it is in this same range of wavelengths that our ability to distinguish one color from another is poorest. A common technique in microscopy uses filters to select just the green light because of the eye's sensitivity, but if detection of color changes is important this is not a good strategy.

Like most of the things that the eye does, the perception of color is determined in a comparative rather than an absolute way. It is only by comparing something to a known color reference that we can really estimate color at all. The usual color reference is a white object, since that (by definition) has all colors. If the scene we are looking at contains something known to be a neutral grey in color, then any variation in the color of the illumination can be compensated for. This is not so simple as it might seem, because many objects do not reflect light of all colors equally and appear to change color with illumination (this "metamerism" is often a problem with ink jet printers).

If the illumination is deficient in some portion of the color spectrum compared to daytime sunlight (which our vision evolved under), then the missing colors can't be reflected or detected and it is impossible to accurately judge the objects true colors. Under monochromatic yellow sodium lights, color is completely confused with albedo (total reflectivity), and we can't distinguish color from brightness. Even with typical indoor lighting, colors appear different (which is why the salesperson suggests you might want to carry that shirt and tie to the window to see how they look in sunlight!).

When Newton first split white sunlight into its component parts using a prism, he was able to show that there were invisible components beyond both ends of the visible spectrum by measuring the rise in temperature that was caused by the absorption of the light. Since sunlight extends well beyond the narrow 400-700 nm range of human vision, it is not surprising that some animals and insects have evolved vision systems that can detect it. Plants, in particular, have developed signals that are visible only in these extended colors in order to attract birds and insects to pollinate them. Extending our human vision into these ranges and even beyond is possible with instrumentation. UV microscopy, radio astronomy, X-ray diffraction all use portions of the electromagnetic spectrum beyond the visible, and all produce data that is typically presented as images, with colors shifted into the narrow portion of the spectrum we can detect.

Being able to detect brightness or color is not the same thing as being able to measure it or detect small variations in either brightness or color. While human vision functions over some 9-10 orders of magnitude, we cannot view a single image that covers such a wide range, nor detect variations of one part in 10^9. A change in brightness of about 2-3% over a lateral distance of a few arc minutes is the limit of detectability under typical viewing conditions. It is important to note that the required variation is a percentage, so that a greater absolute change in brightness is required in bright areas than in dark ones. Anyone with darkroom experience is aware that different details are typically seen in a negative than in a positive image, as shown in Figure 7.

Overall, the eye can detect only about 20-30 shades of grey in an image, and in many cases fewer will produce a visually satisfactory result. In an image dominated by large areas of different brightness, it is difficult to pick out the fine detail with small local contrast within each area. One of the common methods for improving the visibility of local detail is computer enhancement that reduces the global (long-range) variation in brightness while increasing the local contrast (Figure 8). This is typically done by comparing a pixel to its local neighborhood. If the pixel is slightly brighter than its neighbors, it is made brighter still, and vice versa.

Local and abrupt changes in brightness (or color) are the most readily noticed details in images. Variations in texture often represent structural variations that are important, but these are more subtle. As shown in

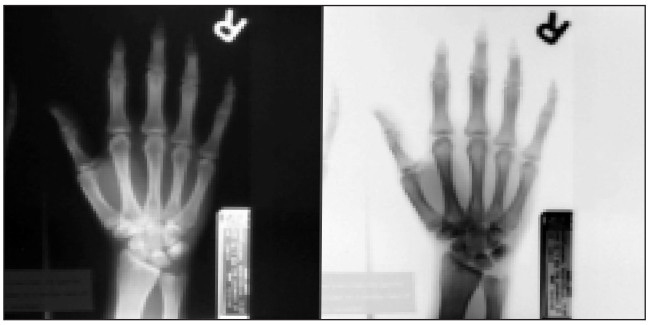


Figure 7. Positive and negative representations of an X-ray. Somewhat different details are visually evident in these images because human vision is not linear, but detects proportional changes in brightness.

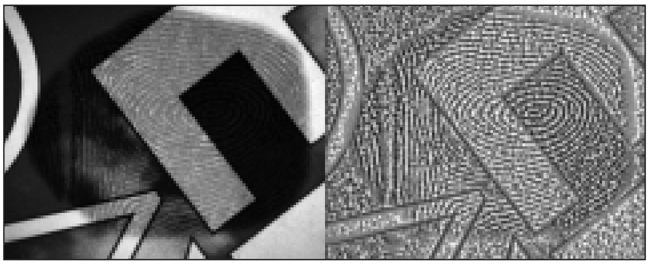


Figure 8. Local contrast enhancement allows visual inspection of low-contrast detail in both the bright and dark regions of this image of a fingerprint on a high contrast magazine cover.

Figure 9, in many cases variations that are classified as textural actually represent changes in the average brightness level. When only a change in texture is present, visual detection is difficult. If the boundaries are not straight lines (and especially vertical or horizontal in orientation), they are much more difficult to see.

Thirty shades of brightness in each of the red, green and blue cones would suggest that $30^3 = 27,000$ colors might be distinguished but that is not so. Sensitivity to color changes at the ends of the spectrum is much better than in the middle (in other words, greens are hard to distinguish from each other). Only about a thousand different colors can be distinguished. Since computer displays offer 256 shades of brightness for the R, G and B phosphors, or $256^3 = 16$ million colors, we might expect that they could produce any color we can see. However, this is not the case. Both computer displays and printed images suffer severe limitations in gamut - the total range of colors that can be produced - as compared to what we can see. This is another reason that the "typical image" may not actually be representative of the object or class of objects.

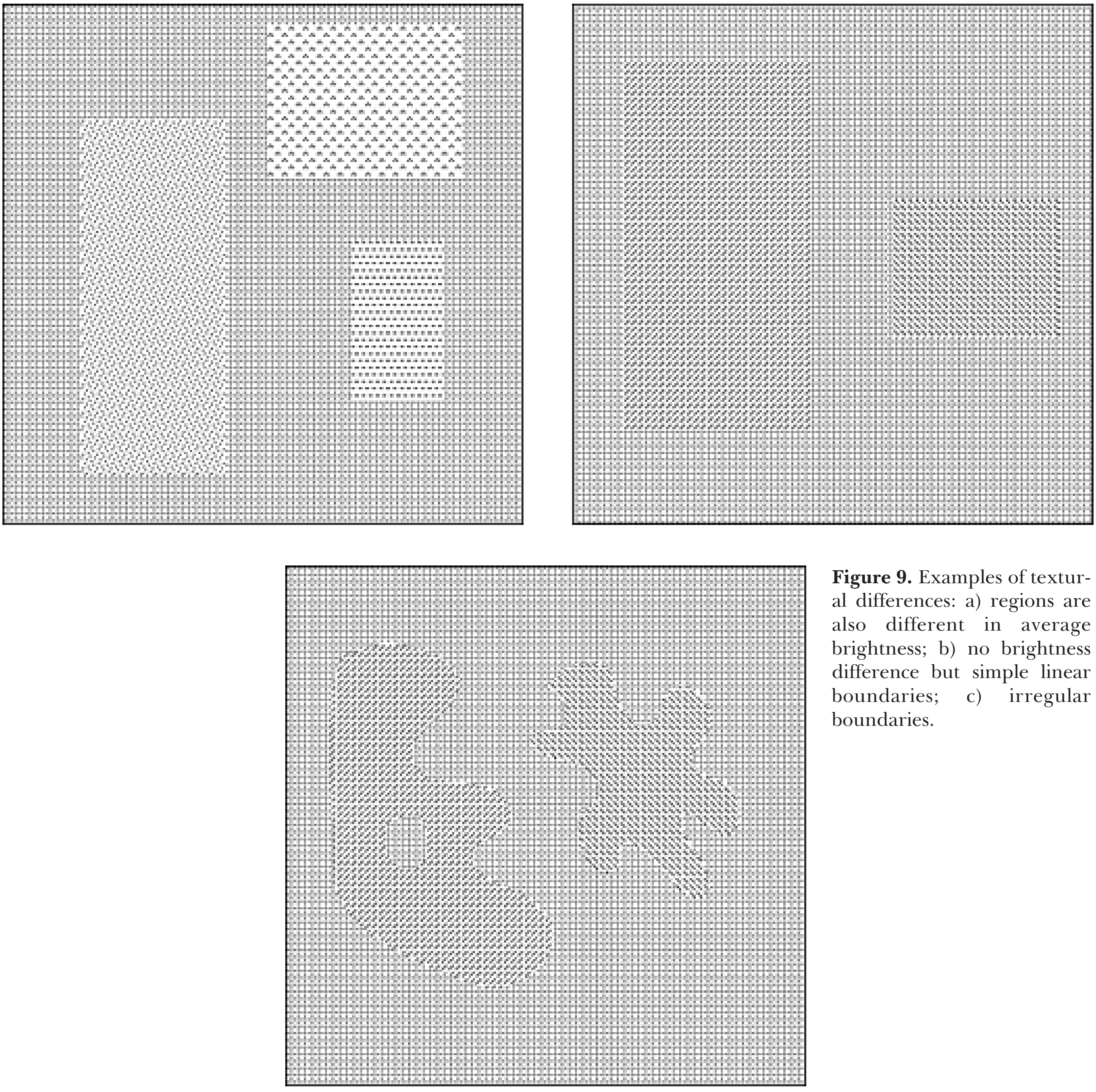


Figure 9. Examples of textural differences: a) regions are also different in average brightness; b) no brightness difference but simple linear boundaries; c) irregular boundaries.

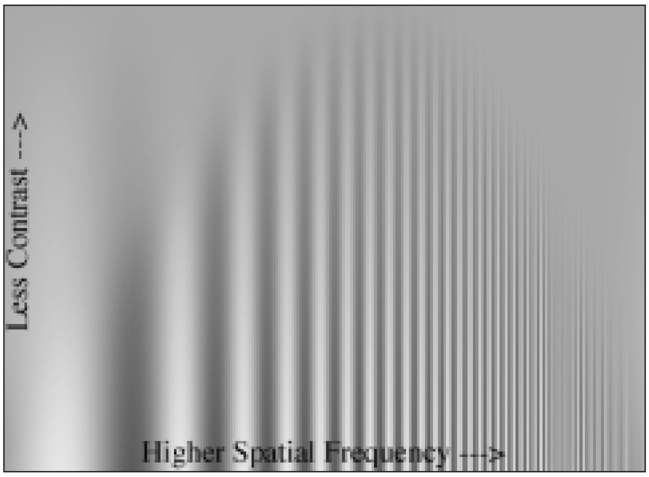


Figure 10. Illustration of the modulation transfer function for human vision, showing the greatest ability to resolve low-contrast details occurs at an intermediate spatial frequency, and becomes poorer for both smaller and larger details.

Acuity

Many animals, particularly birds, have vision that produces much higher spatial resolution than humans. Human vision achieves its highest spatial resolution in just a small area at the center of the field of view (the fovea), where the density of light-sensing cones is highest. At a 50 centimeter viewing distance, details with a width of 1 mm represent an angle of 0.11 degrees. Acuity (spatial resolution) is normally specified in units of cycles per degree. The upper limit (finest detail) visible with the human eye is about 50 cycles per degree, which would correspond to a grating in which the brightness varied from minimum to maximum about 5 time over that same 1 mm. At that fine spacing, 100% contrast would be needed, in other words black lines and white spaces. This is where the common specification arises that the finest lines distinguishable without optical aid are about 100 μm.

Less contrast is needed between the light and dark locations to detect them when the features are larger. Brightness variations about 1 mm wide represent a spatial frequency of about 9 cycles per degree, and under ideal viewing conditions can be resolved with a contrast of less than 1%, although this assumes the absence of any noise in the image and a very bright image (acuity drops significantly in dark images or ones with superimposed random variations, and is much poorer at detecting color differences than brightness variations).

At a normal viewing distance of about 50 cm, 1 mm on the image is about the optimum size for detecting the presence of detail. On a typical computer monitor that corresponds to about 4 pixels. As the spatial frequency drops (features become larger) the required contrast increases, so that when the distance over which the brightness varies from minimum to maximum is about 1 cm, the required contrast is about 10 times greater. The variation of spatial resolution ("acuity") with contrast is called the modulation transfer function (Figure 10).

Enlarging images does not improve the ability to distinguish small detail, and in fact degrades it. The common mistake made by microscopists is to work at very high magnification expecting to see the finest details. That may be needed for details that are small in dimension, but it will make it more difficult to see larger features that have less contrast.

Because the eye does not "measure" brightness, but simply makes comparisons, it is very difficult to distinguish brightness differences unless the regions are immediately adjacent. Figure 11a shows four grey squares, two of which are 5% darker than the others. Because they are separated, the ability to compare them is limited. Even if the regions are adjacent, as in Figure 11b, if the change from one region to another is gradual it cannot be detected. Only when the step is abrupt, as in Figure 11c, can the eye easily determine which regions are different.

Even when the features have sizes and contrast that should be visible, the presence of variations in the background intensity (or color) can prevent the visual system from detecting them. In the example of

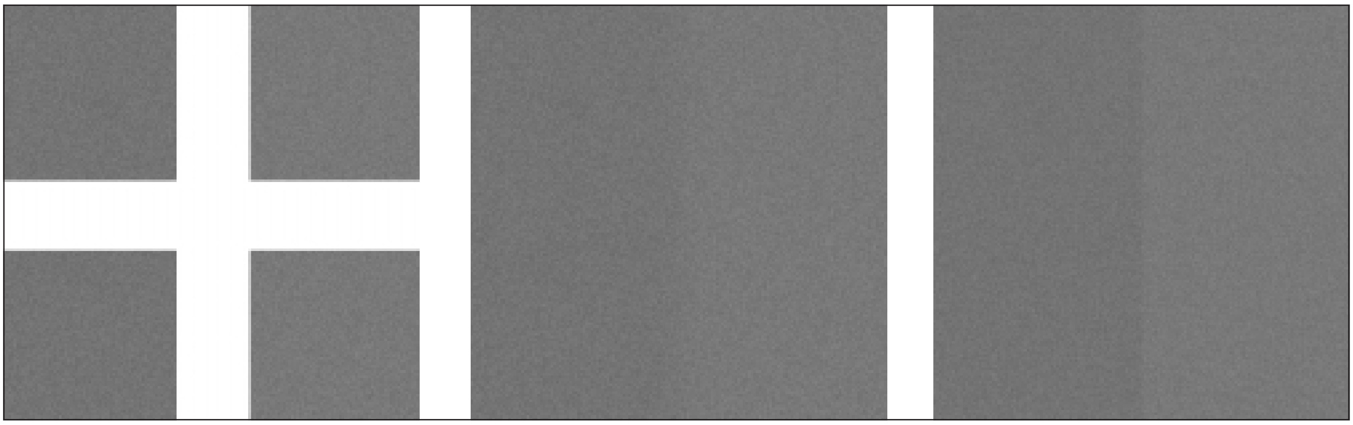


Figure 11. Comparison of regions with a 5% brightness difference: a) separated; b) adjacent but with a gradual change; c) adjacent with an abrupt boundary.

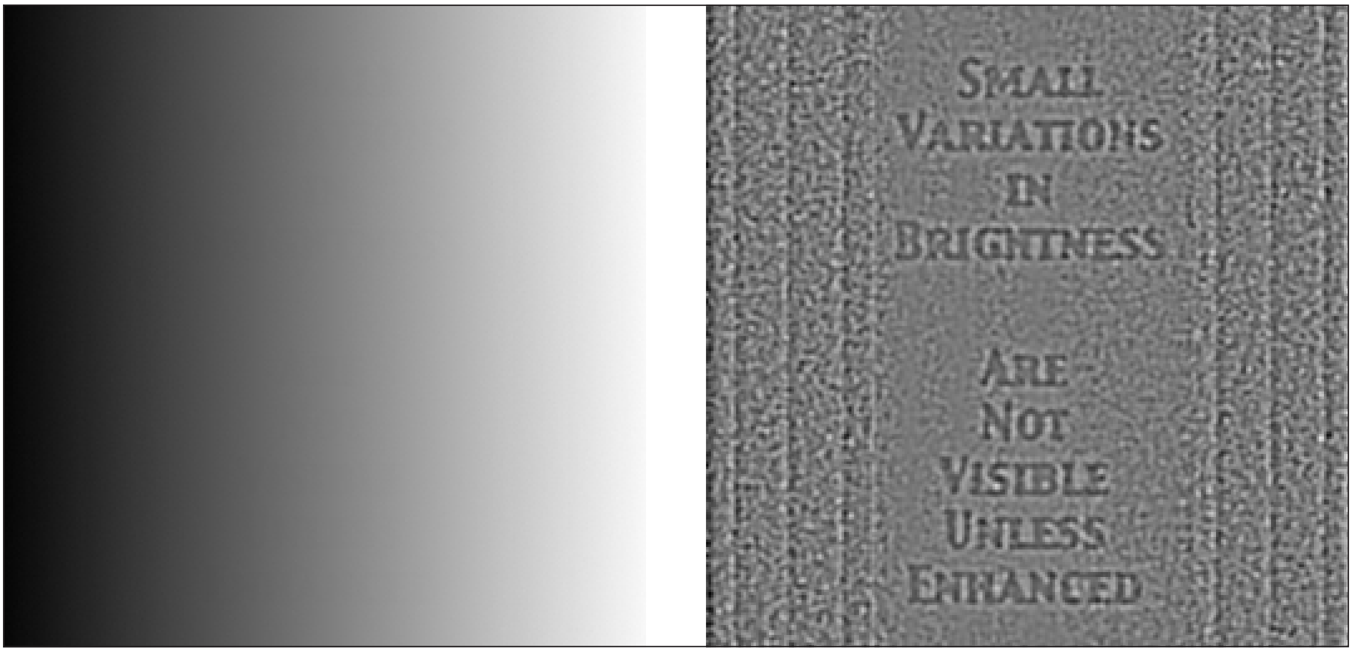
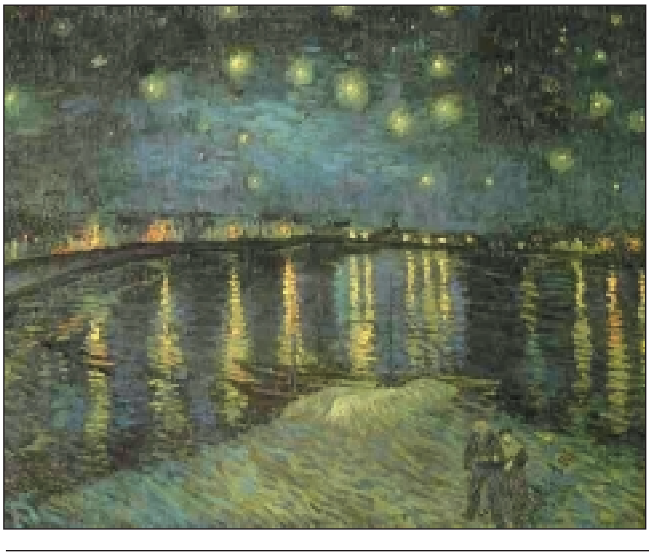


Figure 12. Intensity differences superimposed on a varying background are visually undetectable: a) original; b) processed with an "Unsharp Mask" filter to suppress the gradual changes and reveal the detail.

Figure 12, the text has a local contast of about 1% but is superimposed on a ramp that varies from white to black. Application of an image processing operation reveals the message. The "Unsharp Mask" routine subtracts a blurred (smoothed) copy of the image from the original, suppressing large scale variations in order to show local details.

It is easy to confuse resolution with visibility. A star in the sky is essentially a point; there is no angular size, and even in a telescope it does not appear as a disk. It is visible because of its contrast, appearing bright against the dark sky. Faint stars are not visible to the naked eye because there isn't enough contrast. Telescopes make them visible by collecting more light into a larger aperture. A better way to think about resolution is the ability to distinguish as separate two stars that are close together. The classic test for this has long been the star Mizar in the handle of the big dipper. In Van Gogh's "Starry Night over the Rhone" each of the stars in this familiar constellation is shown as a single entity (Figure 13). But a proper star chart shows that the second star in the handle is actually double. Alcor and Mizar are an optical double - two stars that appear close together but in fact are at different distances and have no gravitational relationship to each other. They are separated by about 11.8 minutes of arc, and being able to detect the two as separate has been considered by many cultures from the American Indians to the desert dwellers on the near east as a test of good eyesight (as well as the need for a dark sky with little turbulence or water vapor, and without a moon or other light pollution).

But there is more to Mizar than meets the eye, literally. With the advent of the Galilean telescope, observers were surprised to find that Mizar itself is a double star. Giovanni Battista Riccioli (1598 - 1671), the Jesuit astronomer and geographer of Bologna, is generally supposed to have split Mizar, the first double star ever discovered, around 1650. The two stars Mizar-A and Mizar-B, are a gravitational double 14.42



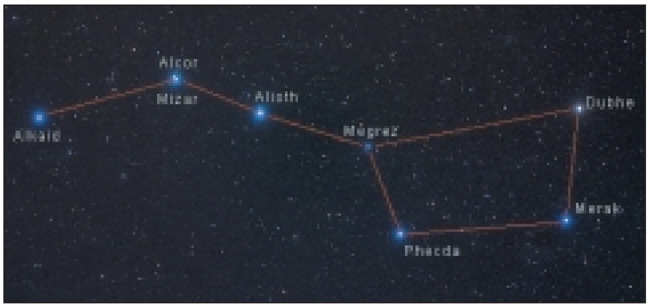


Figure 13. The big dipper, in Van Gogh's "Starry Night Over the Rhone", and as shown in a star chart.

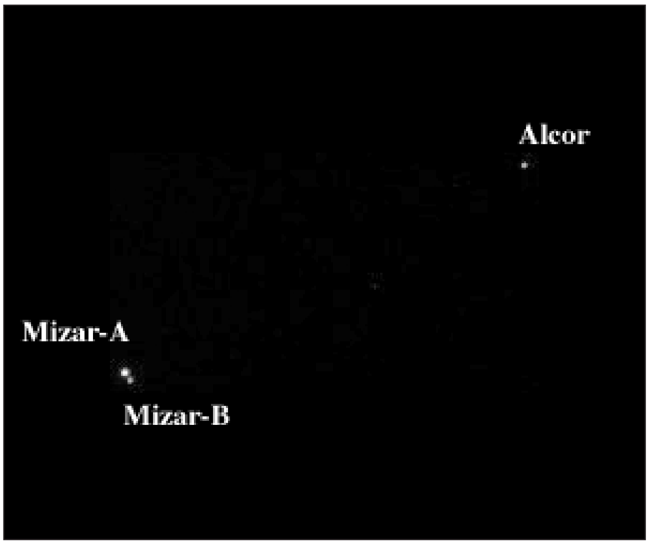


Figure 14. Photograph showing the relative separation of Alcor from Mizar A and B.

arc seconds apart, and any good modern telescope can separate them (Figure 14). But they turn out to be even more complicated - each star is itself a double as well, so close together that only spectroscopy can detect them.

There are many double stars in the sky that can be resolved with a backyard telescope, and many of them are familiar viewing targets for amateur astronomers who enjoy the color differences between some of the star pairs, or the challenge of resolving them. This depends on more than just the angular separation. If one star is significantly brighter than the other, it is much more difficult to see the weaker star close to the brighter one.

What the eye tells the brain

Human vision is a lot more than rods and cones in the retina. An enormous amount of processing takes place, some of it immediately in the retina and some in the visual cortex at the rear of the brain, before an "image" is available to the conscious mind. The neural connections within the retina were first seen about a hundred years ago by Ramón y Cajal, and have been studied ever since. Figure 15 shows a simplified diagram of the human retina. The light-sensing rods and cones are actually at the back, and light must pass through several layers of processing cells to reach them. In many nocturnal animals the pigmented layer behind the rods and cones reflects light back so that the photons are twice as likely to be captured and detected (and some comes back out through the lens to produce the reflective eyes we see watching us at night). Incidentally, the eye of the octopus does not have this backwards arrangement; evolution in that case put the light sensing cells on top where they can most efficiently catch the light.

The first layer of processing cells, called horizontal cells, connect light sensing cells in various size neighborhoods. The next layer, the amacrine cells, combine and compare the outputs from the horizontal cells. Finally the ganglion cells collect the outputs for transmission to the visual cortex. This physical organization corresponds directly to the logical processes of inhibition, discussed below.

In many respects, the retina of the eye is actually part of the brain. Understanding the early processing of image data and the extraction of the information transmitted from the eye to the visual cortex has been awarded the Nobel prize (in 1981, to David H. Hubel and Torsten N. Wiesel, for their discoveries concerning information processing in the visual system).

Without elaborating the anatomical details of the retina or the visual cortex discovered by Hubel, Wiesel, and others (particularly a seminal paper "What the frog's eye tells the frog's brain" published in 1959 by Jerome Lettvin), it is still possible to sum up the logical and practical implications. Within the retina, outputs from the individual light sensors are combined and compared by layers of neurons. Comparing the output from one sensor or region to that from the surrounding sensors, so that excitation of the center is tested against the inhibition from the surroundings, is a basic step that enables the retina to ignore regions that are uniform or only gradually varying in brightness, and to efficiently detect locations where a change in brightness occurs. Testing over different size regions locates points and features of varying sizes. Comparison of output over time is carried out in the same way to detect changes.

In the frog's eye, the retina processes images to find just a few highly specific stimuli. These include small dark moving objects (food) and the location of the largest, darkest region (safety in the pond). Like the fly and a human, the eye is hard-wired to detect "looming," something that grows rapidly and does not shift in the visual field. That represents something coming toward the eye, and causes the fly to avoid the

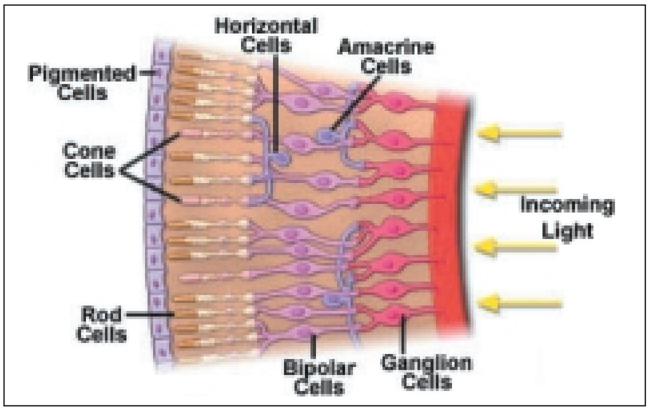


Figure 15. The principal layers in the retina. Light passes through several layers of processing neurons to reach the light-sensitive rod and cone cells. the horizontal, bipolar and amacrine cells combine the signals from various size regions, compare them to locate interesting features, and pass that information on to higher levels in the visual cortex.

swatter or the frog to jump into the water. In a human, the eyelid blinks for protection without our ever becoming consciously aware that an object was even seen.

The outputs from these primitive detection circuits are then further combined in the cortex to locate lines and edges. There are specific regions in the cortex that are sensitive to different orientations of lines and to their motion. The "wiring" of these regions is not built in, but must be developed after birth; cats raised in an environment devoid of lines in a specific orientation do not subsequently have the ability to see such lines. Detection of the location of brightness changes (feature edges) creates a kind of mental sketch of the scene, which is dominated by the presence of lines, edges, corners and other simple structures. These in turn are linked together in a hierarchy to produce the understanding of the scene in our minds.

The extraction of changes in brightness or color with position or with time explains a great deal about what we see in scenes, and about what we miss. Changes that occur gradually with position, such as shading of light on a wall, is ignored. We have to exert a really conscious effort to notice such shading, and even then we have no quantitative tools with which to estimate its magnitude. But even small changes in brightness of a few percent are visible when they occur abruptly, producing a definite edge. And when that edge forms a straight line (and particularly when it is vertical) it is noticed. Similarly, any part of a scene that is static over time tends to be ignored, but when something moves it attracts our attention.

These techniques for extracting information from scenes are efficient because they are highly parallel. For every one of the 160 million light sensors in the eye, there are as many as 50,000 neurons involved in processing and comparing. One of Hubel and Wiesel's contributions was showing how the connections in the network are formed shortly after birth, and the dependence of that formation on providing imagery to the eyes during that critical period.

Mapping of the specific circuitry in the brain is accomplished by placing electrodes in various locations in the cortex, and observing the output of neurons as various images and stimuli are presented to the eyes. At a higher level of scale and processing, functional MRI and PET scans can identify regions of the brain that respond to various activities and stimuli. But there is another source of important knowledge about processing of images in the mind: identifying the mistakes that are made in image interpretation. One important, but limited resource is studying the responses of persons who have known specific damage, either congenital or as the result of illness or accident. A second approach studies the errors in interpretation of images resulting from visual illusions. Since everyone tends to make the same mistakes, those errors must be a direct indication of how the processing is accomplished. Several of the more revealing cases will be presented in the sections that follow.

Spatial comparisons

The basic idea behind center-surround or excitation-inhibition logic is comparing the signals from a central region (which may be a single detector, or progressively larger scales by averaging detectors together) to the output from a surrounding annular region. That is the basic analysis unit in the retina, and by combining the outputs from many such primitive units, the detection of light or dark lines, corners, edges and other structures can be achieved. In the frog's eye, it was determined that a dark object of a certain size (corresponding to an insect close enough to be captured) generated a powerful recognition signal. In the cat's visual cortex there are regions that respond only to dark lines of a certain length and angle, moving in a particular direction. Furthermore, those regions are intercalated with regions that perform the same analysis on the image data from the other eye, which is presumed to be important in stereopsis or fusion of stereo images.

This fundamental center-surround behavior explains several very common illusions (Figure 16). A set of uniform grey steps (Mach bands) are not perceived as being uniform. The side of each step next to a lighter region is perceived as being darker, and vice versa,

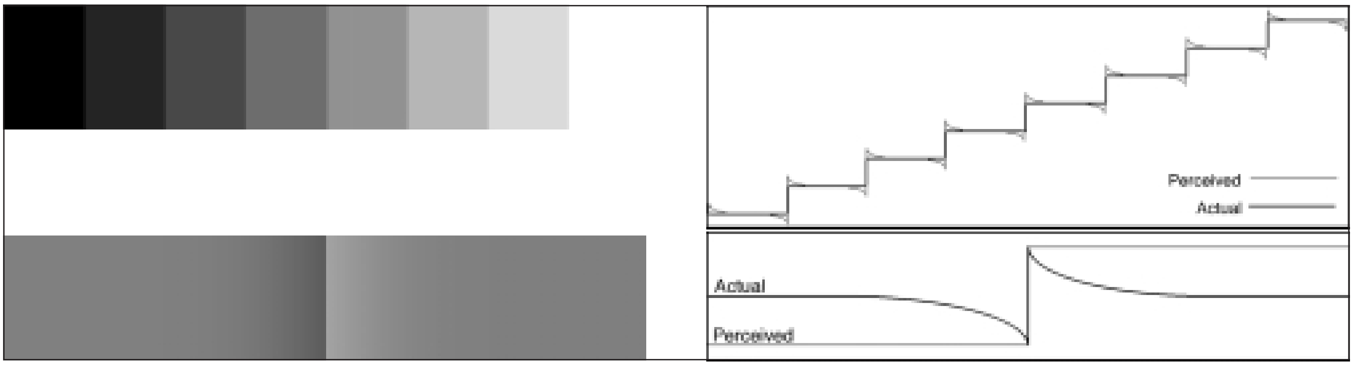


Figure 16. Two common illusions based on inhibition: Mach bands (top) demonstrate that the visual system increases the perceived change in brightness at steps; the Craik-Cornsweet-O'Brien step (bottom) shows that providing the visual system with a step influences the judgment of values farther away.

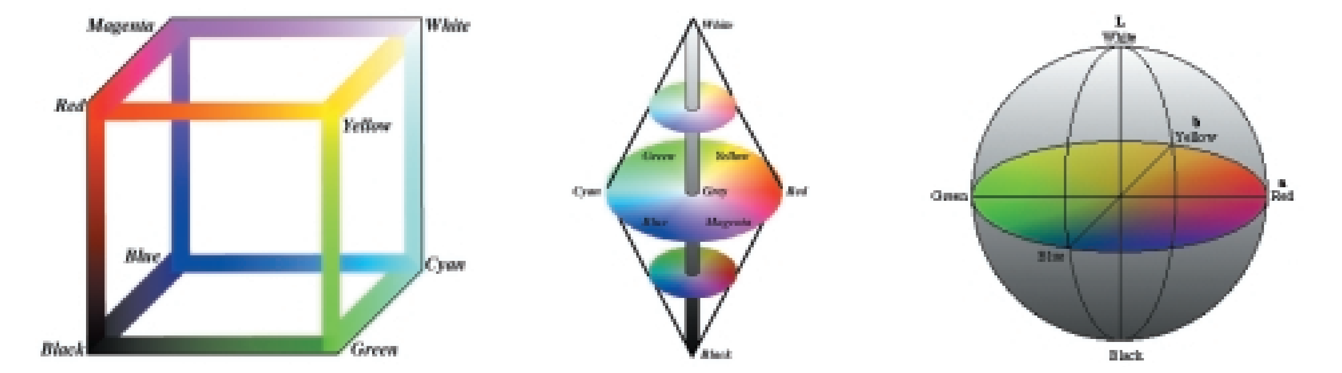


Figure 17. Diagrams of color spaces: a) the RGB color space is a cube that is mathematically simple and describes the way cameras and displays work, but not the way people think about color; b) HSI space has a central intensity (also called brightness, value or luminance) axis that is pure grey scale, while the distance out from the axis (saturation) measures the amount of color and the angle (hue) identifies what the color is (the Hue-Saturation plane is the color wheel familiar to school children); c) Lab space is a sphere with a vertical axis for Luminance (brightness) and to perpendicular color axes, one for red-green and the other for blue-yellow (this is mathematically simpler than HSI and still separates intensity from color information).

because it is the step that is noticed. In the case of the Mach bands, the visual response to the step aids us in determining which region is darker, although it makes it very difficult to judge the amount of the difference. But the same effect can be used in the absence of any difference to cause one to be perceived. In the Craik-Cornsweet-O'Brien illusion, two adjacent regions have the same brightness but the values are raised and lowered on either side of the boundary. The eye interprets this in the same way as for the Mach bands, and judges that one region is, in fact, lighter and one darker.

Similar excitation-inhibition comparisons are made for color values. Boundaries between blocks of color are detected and emphasized, while the absolute differences between the blocks are minimized. Blocks of identical color placed on a gradient of brightness or some other color will appear to be different. The human response to color is not the same as a camera's (either digital or film). Although the cones in the retina respond to short (blue), medium (green) and long (red) wavelengths, combinations and comparisons performed early in the visual process convert this to a different "space" for color interpretation.

Whether described in terms of brightness, hue and saturation or the artist's tint, shade and tone, or various mathematical spaces such as Ycc (used in color video) or Lab (used in many image processing software routines), three parameters are needed (Figure 17). Brightness is a measure of how light or dark the light is, without regard to any color information. For the artist, this is the shade, achieved by adding black to the pigment. Hue distinguishes the various colors, progressing from red to orange, yellow, green, blue, and magenta around the color wheel we probably encountered in Kindergarten (or the ROY G BIV mnemonic for the order of colors in the rainbow). Hue corresponds to the artist's pure pigment. Saturation is the amount of color present, such as the difference between pink and red, or sky blue and royal blue. A fully saturated color corresponds to the artist's pure pigment, which can be tinted with white pigment to reduce the saturation (tone describes adding both white and black pigments to the pure color).

Even without formal training in color science, this is how people describe color to themselves. Combination and comparison of the signals from the red, green and blue sensitive cones is used to determine these parameters. Simplistically, we can think of the sum of all three being the brightness, ratios of one to another being interpreted as hue, and the ratio of the greatest to the least giving the saturation. One important thing to remember about hue is that it does not correspond to the linear wavelength range from red to blue. We interpret a color with reduced green but strong red and blue as being magenta, which is not a color in the wavelength sense but certainly is in terms of perception.

Relying on comparisons to detect changes in brightness or color rather than absolute values simplifies many tasks of image interpretation. For example, the various walls of a building, all painted the same color, will have very different brightness values because of shading, their angle to the sun, etc., but what we observe is a building of uniform color with well defined corners and detail because it is only the local changes that count. The "white" walls on the shadowed side of the building in Figure 18 are actually darker than the "black" shingles on the sunlit side, but our perception understands that the walls are white and the shingles black on all sides of the building.

The same principle of local comparisons applies to many other situations. Many artists have used something like the Craik-Cornsweet-O'Brien illusion to create the impression of a great range of light and shadow within the rather limited range of absolute brightness values that can be achieved with paint on canvas (Figure 19). The effect can be heightened by

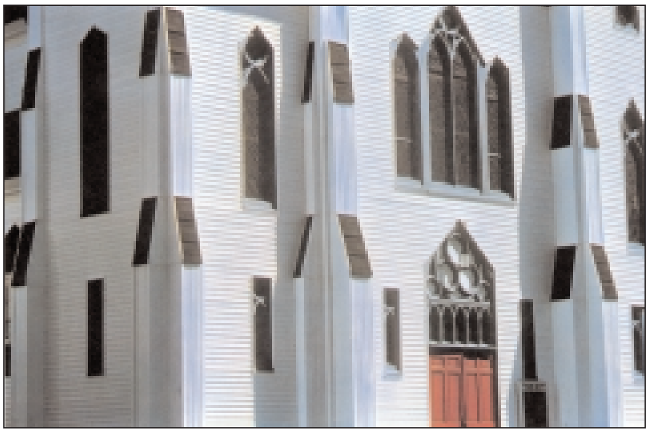


Figure 18. Comparison of the dark and light areas on the shadowed and sunlit sides of the building is performed locally as discussed in the text.

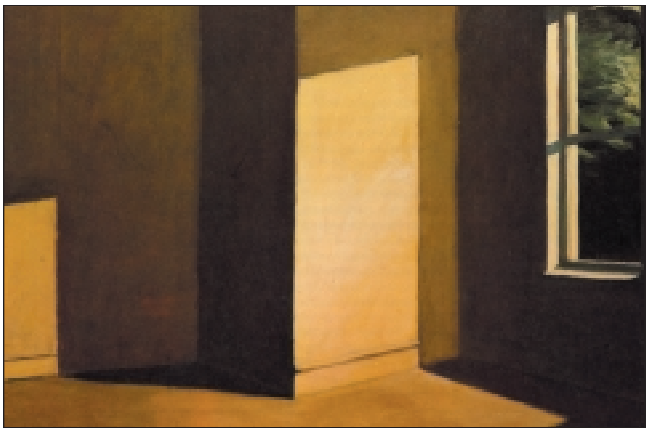


Figure 19. Edward Hopper painted many renderings of light and shadow. His "Sunlight in an Empty Room" illustrates shading along uniform surfaces and steps at edges.

adding colors, and there are many cases in which shadows on skin are tinted with colors such as green or magenta to increase the perception of a brightness difference.

Edwin Land (of Polaroid fame) studied color vision extensively and proposed a somewhat different interpretation than the tri-stimulus model described above (which is usually credited to Helmholtz). Also relying heavily on the excitation-inhibition model for processing signals, Land's opponent-color "retinex" theory predicts or explains several interesting visual phenomena. It is important to note that in Land's retinex theory the composition of light from one region in an image considered by itself does not specify the perceived color of that area, but rather that the color of an area is determined by a trio of numbers, each computed on a single waveband (roughly the long, medium and short wavelengths usually described as red, green and blue) to give the relationship on that waveband between that area and the rest of the areas in the scene.

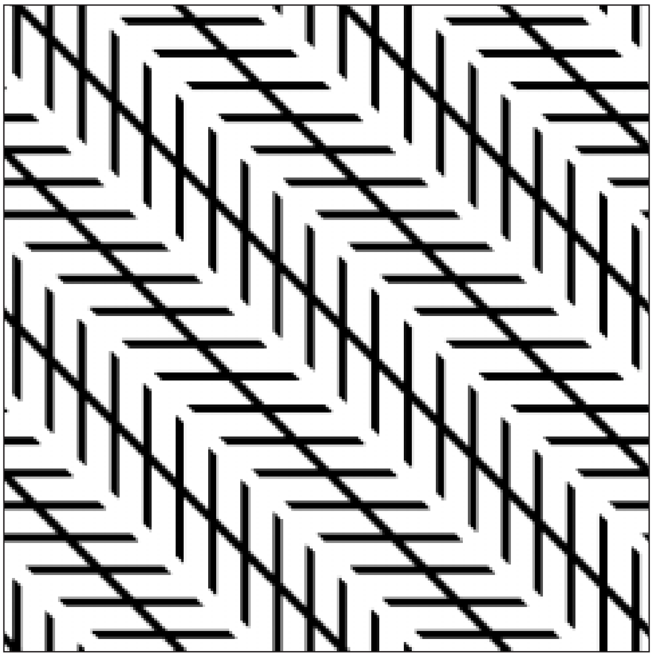


Figure 20. Zollner lines. The cross-hatching of the diagonal lines with short vertical and horizontal ones causes our visual perception of them to rotate in the opposite direction. In fact, the diagonal lines are exactly parallel.

One of the consequences of Land's theory is that the spectral composition of the illumination becomes very much less important, and the color of a particular region can become relatively independent of the light that illuminates it. Our eyes have relatively few cones to sense the colors in the periphery of our vision, but that information is apparently very important in judging colors in the center by correcting for the color of incident illumination. Land demonstrated that if a scene is photographed through red, green and blue filters, and then projectors are used to shine colored and/or white light through just two of the negatives, the scene may be perceived as having full color. Another interesting phenomenon is that a spinning disk with black and white bars may be perceived as having color, depending on the spacing of the bars. Many of the questions about exactly how color information is processed in the human visual system have not yet been answered.

The discussion of center-surround comparisons above primarily focused on the processes in the retina, which compare each point to its surroundings. But as the extracted information moves up the processing chain, through the visual cortex, there are many other evidences of local comparisons. It was mentioned that in the visual cortex there are regions that respond to lines of a particular angle. They are located adjacent to regions that respond to lines of a slightly different angle. Comparison of the output from one region to its neighbor can thus be used to detect small changes in angle, and indeed that comparison is carried out.

We don't measure angles with mental protractors, but we do compare them and notice differences. Like the case for brightness variations across a step, a difference in angle is amplified to detect boundaries and increase the perceived change. In the absence of any markings on the dial of a wristwatch, telling time to the nearest minute is about the best we can do. One minute corresponds to a six degree motion of the minute hand. But if there are two sets of lines that

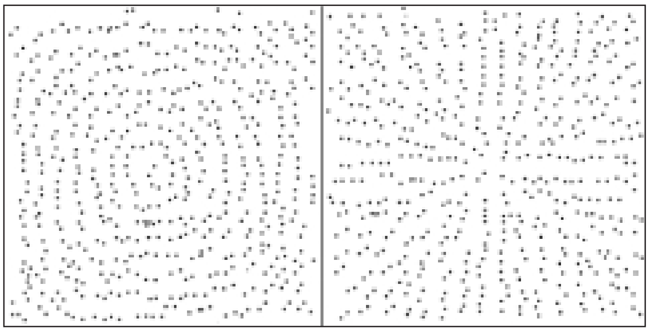


Figure 21. Connecting each point to its nearest neighbor produces radial or circumferential lines.

vary by only a few degrees, we notice that difference (and usually judge it to be much larger).

Inhibition as regards to angle means that cross-hatching diagonal lines with short marks that are vertical or horizontal will alter our perception of the main line orientation. As shown in Figure 20, this makes it difficult or impossible to correctly compare the angle of the principal lines (which are in fact parallel). Human vision does not treat all angles the same, because we did not evolve in a world where all directions are the same. Very small deviations from vertical, and somewhat larger ones from horizontal, are readily detected, while very large changes are required to be detected at arbitrary diagonal orientations.

Local to global hierarchies

Interpretation of elements in a scene relies heavily on grouping them together to form features and objects. At very simple example (Figure 21) using just dots shows that the eye connects nearest neighbors to construct alignments.

Similar grouping of points and lines is used to connect together parts of feature boundaries that are otherwise poorly defined, to create the outlines of objects that we visually interpret (Figure 22).

It is a simple extension of this grouping to include points nearest over time that produces the familiar "Star Trek" impression of motion through a starfield (Figure 23). Temporal comparisons are discussed more fully below.

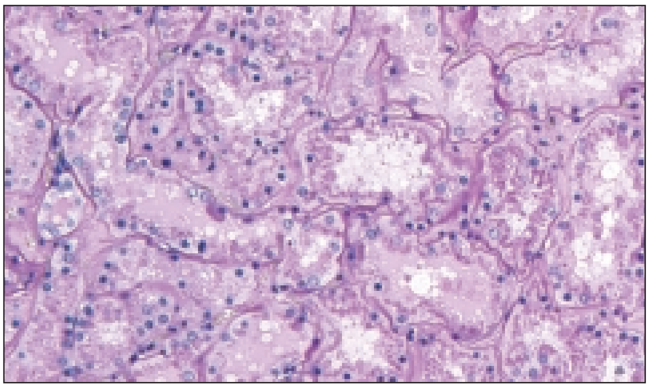


Figure 22: The cell boundaries in this tissue section are visually constructed by grouping the lines and points.

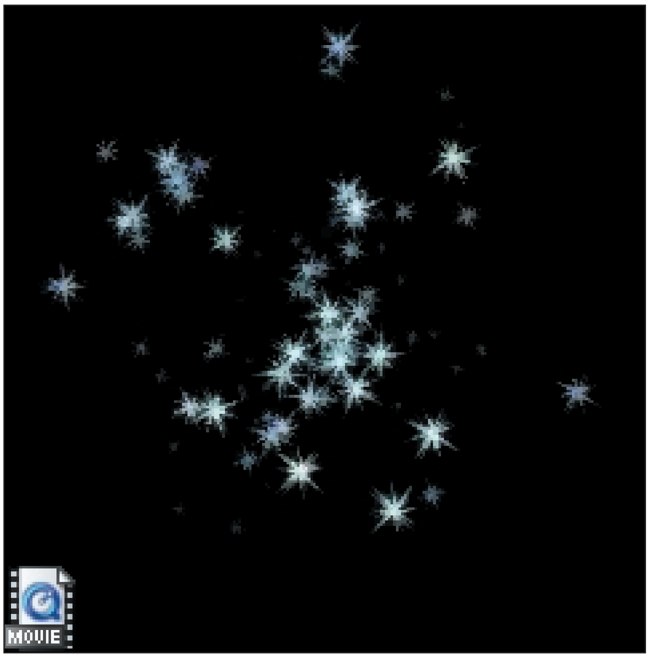


Figure 23. Star Trek introduced the "moving starfield" graphic (this Quicktime movie can be downloaded from http://DrJohnRuss.com/images/Seeing/Fig_23.mov).

Our natural world does not consist of lines and points, but of the objects that they connect together to represent. Our vision systems perform this grouping naturally, and it is usually difficult to deconstruct a scene or structure into its component parts. Again, illusions are very useful to illustrate the way this grouping works (Figure 24). The lengths of two simple parallel lines in isolation are easily compared. But if additional lines are connected to these, they become visually part of the same structure and the perceived lengths of the original lines are altered. In the arrow illusion, the different orientation of the added lines does not separate them from the main shaft. But if they are very different in color, the illusion weakens or fails because they are not grouped together.

Grouping is necessary for inhibition to work effectively. The cross-hatching in Figure 20 shifts the angle of the line because it is seen as being part of it. The com-

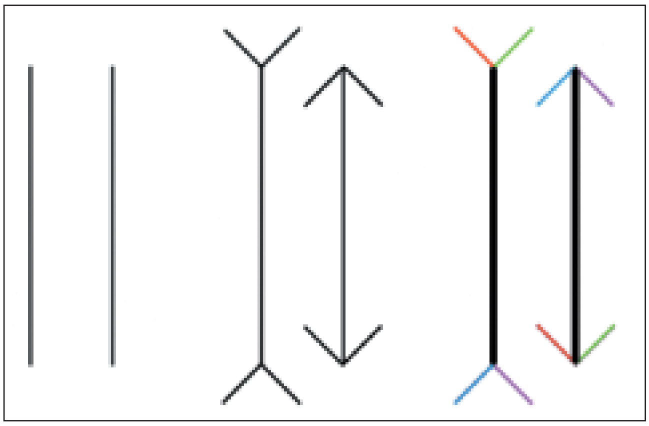


Figure 24. Example of grouping. The slanted lines added to the ends of the (identical) horizontal lines affects our visual comparison of their length. This effect is reduced when they are different in color or thickness.

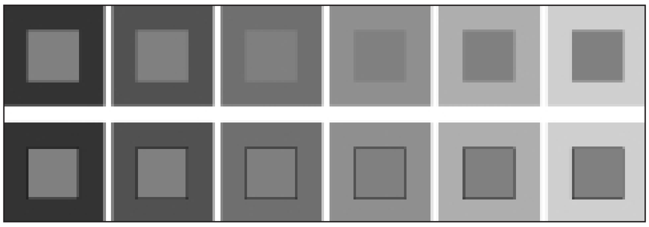


Figure 25. In the top image the lighter and darker surrounds affects our visual comparison of the (identical) central grey regions. This effect is reduced in the bottom image by the separating border.

mon illusion of brightness alteration due to a surrounding, contrasting frame depends on the frame being grouped with the central region. In the example shown in Figure 25, the insertion of a black line separating the frame from the center alters the appearance of the image and reduces or eliminates the perceived difference in brightness of the grey central regions. Without the line, the brighter and darker frames are grouped with the center and inhibition alters the apparent brightness, making the region with the dark frame appear lighter and vice versa.

The combination of grouping with inhibition gives rise to the illusion shown in Figure 26. The angled shading causes the lines to appear slightly turned in the opposite direction, and the various pieces of line are connected by grouping, so that the entire figure is perceived as a spiral. But in fact, the lines are circles, as can be verified by tracing around one of them. This is an example of spatial grouping, but temporal grouping works in much the same way. A rotating spiral is commonly seen as producing an endless motion, whether it is a barber pole or the spinning disk used in every cheap science fiction movie ever made (Figure 27).

Grouping together features in an image that are the same in color lies at the very heart of the common tests for color blindness. In the Ishihara tests, a set of colored circles are arranged so that similar colors can be grouped to form recognizable numbers (Figure 28). For the person who cannot differentiate the colors, the same grouping does not occur and the interpretation of the numbers changes. The example shown is one of a set that diagnoses red-green deficiency, the most common form of color blindness that affects perhaps 1 in 10 men. This deficiency can be classed as either protanopia or deuteranopia. In protanopia, the visible range of the spectrum is shorter at the red end compared with that of the normal, and that part of the spectrum that appears blue-green in the normal appears to those with protanopia as grey. In deuteranopia the part of the spectrum that appears to the normal as green appears as grey. Purple-red (the complimentary color of green) also appears as grey. In the example, those with normal color vision should read the number 74. Red-green

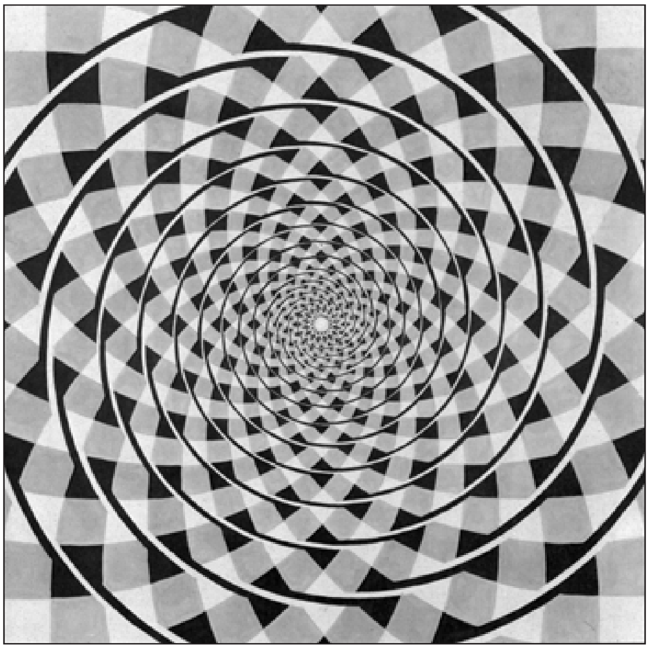


Figure 26. Fraser's spiral. The circular rings are visually "tilted" and perceived to form a spiral.

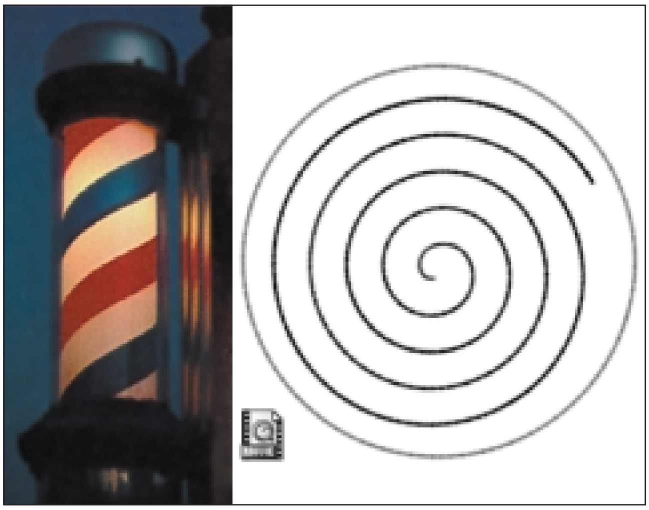


Figure 27. Moving helical and spiral patterns such as the barber pole and spinning disk (this Quicktime movie can be downloaded from http://DrJohnRuss.com/images/Seeing/Fig_27.mov) produce an illusion of motion of the background perpendicular to the lines.

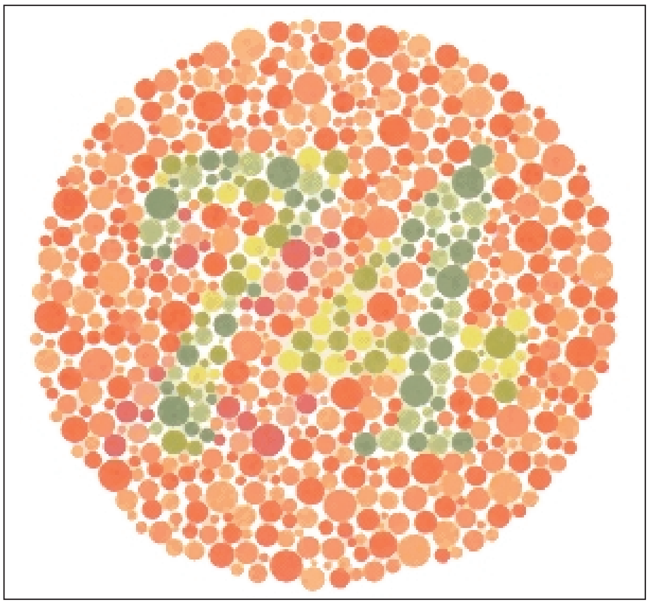


Figure 28. One of the Ishihara color blindness test images (see text)

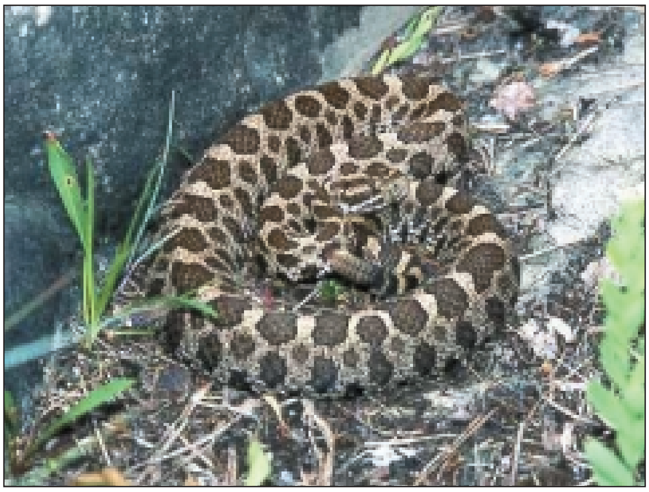


Figure 29. Natural camouflage (a pygmy rattlesnake hides very well on a bed of leaves).

color deficiency will cause the number to be read as 21. Someone with total color blindness will not be able to read any numeral.

By interfering with our visual ability to group parts of the image together, camouflage is used to hide objects. The military discovered this before the first world war, when the traditional white of US Navy ships (Teddy Roosevelt's "Great White Fleet") was replaced by irregular patterns of greys and blues to reduce their visibility. But nature figured it out long ago, and examples are plentiful. Predators wanting to hide from their prey, and vice versa, typically use camouflage as a first line of defense. There are other visual possibilities of course - butterflies whose spots look like the huge eyes of a larger animal (mimicry), or frogs whose bright colors warn of their poison - but usually (as in Figure 29) the goal is simply to disappear. Breaking the image up so that the brain does not group the parts together very effectively prevents recognition.

Motion can destroy the illusion produced by camouflage, because moving objects (or portions of objects) attract notice, and if several image segments are observed to move in coordinated ways they are grouped, and the hidden object emerges. Human vision attempts to deal with moving features or points as rigid bodies, and easily connects separated pieces that move in a coordinated fashion. Viewing scenes or images with altered illumination, or a colored filter, often reveals the objects as well. But in nature, the ability to keep still and rely on camouflage is a well-developed strategy.

Grouping operates on many spatial (and temporal) scales. In a typical scene there may be lines, edges, and other features that group together to be perceived as an object, but then that object will be grouped with others to form a higher level of organization, and so on. Violations that occur in the grouping hierarchy give rise to conflicts that the mind must resolve. Sometimes this is done by seeing only one interpretation and ignoring another, and sometimes the mind switches back and forth between interpretations. Figure 30 shows two examples. If the bright object is perceived as the foreground, it is seen as a vase with an irregular shape. If the background around the vase becomes the foreground, it emerges as two facing human profiles. An even simpler example is the drawing of a cube. The highlighted corner can be seen either as the point closest to the viewer or farthest away.

Figure 31 shows two more familiar examples of rivalrous interpretations. Is the first drawing the head of a bunny rabbit or a duck? Is the second one a young girl or an old woman? For all of these images in which multiple interpretations or foreground-background reversal produce different perceptions, some people

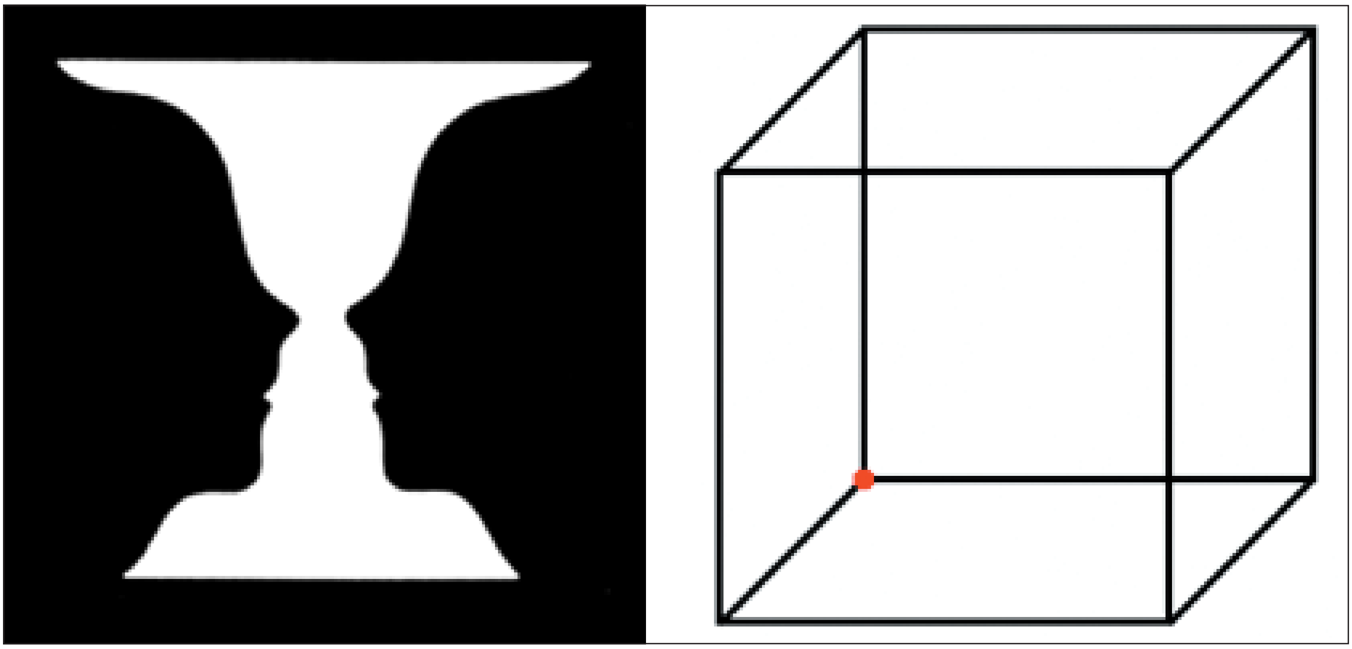


Figure 30. Illusions with two alternative interpretations: a) either two facing profiles or a vase; the Necker cube - is the corner marked in red closest to or farthest from the viewer?

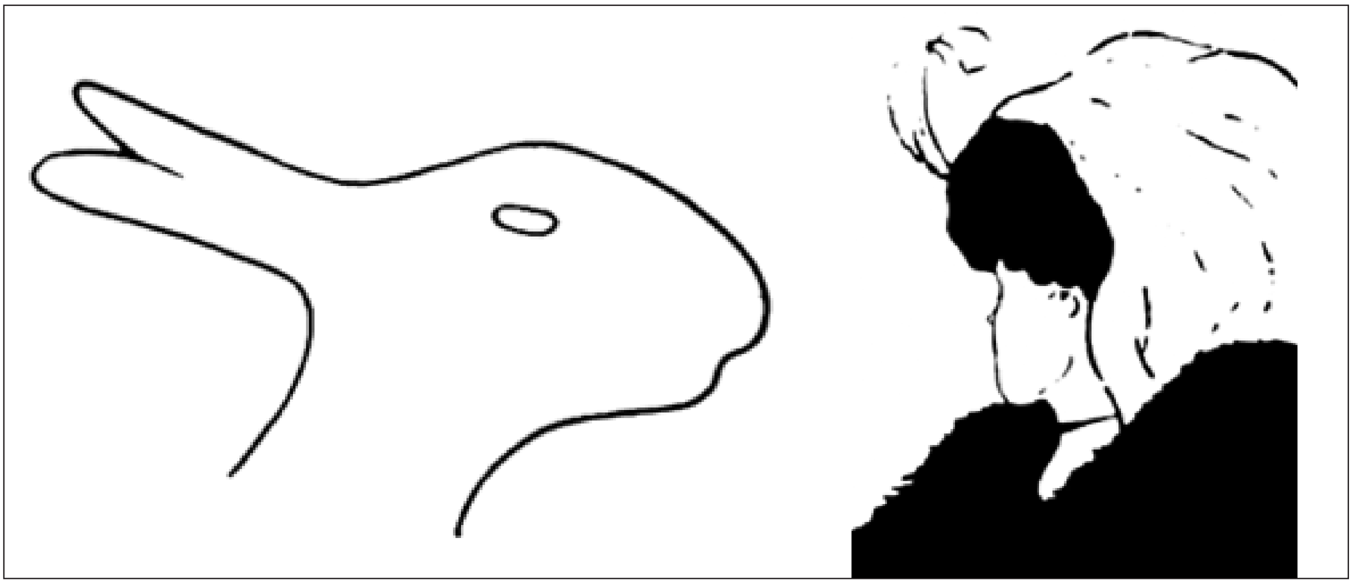


Figure 31. Illusions with alternative interpretations: a) a duck or a rabbit; b) a young girl or an old woman.

initially see only one or the other possibility, and may have difficulty in recognizing the alternative. But once you manage to "see" both interpretations, it is impossible to see both at once, and for most people the two possibilities alternate every few seconds as the image is viewed.

As a trick that may provide some insight into how images are processed in the mind, these examples have some interest. But they also provide a warning for anyone who relies on visual inspection of images to obtain information about the subject. In order to see past camouflage to connect the disparate parts of an object and recognize its presence, you must have a good stored model of what the object is. But when you approach an image with a strong model and try to connect the parts of the image to it, you are forced to ignore other interpretations. Anything in the image that does not conform to that model is likely to be ignored.

In one of Tony Hillerman's detective stories, his Navajo detective Joe Leaphorn explains how he looks for tracks. The FBI man asks "What were you looking for?" and Leaphorn replies "Nothing in particular. You're not really looking for anything in particular. If you do that, you don't see things you're not looking for." But learning how to look for everything and nothing is a hard skill to master. Most of us see only the things we expect to see.

The artist M. C. Escher created many drawings in which grouping hierarchy was consistent locally but not globally, to create conflicts and rivalries that make the art very interesting. Figure 32 shows diagrams and examples of some of his simpler conundrums. In one, lines that represent one kind of edge at the tip of the fork become something different at its base, exchanging inside for outside along the way. In the second, the front-to-rear order of edges changes. Edges that occlude others and are therefore per-

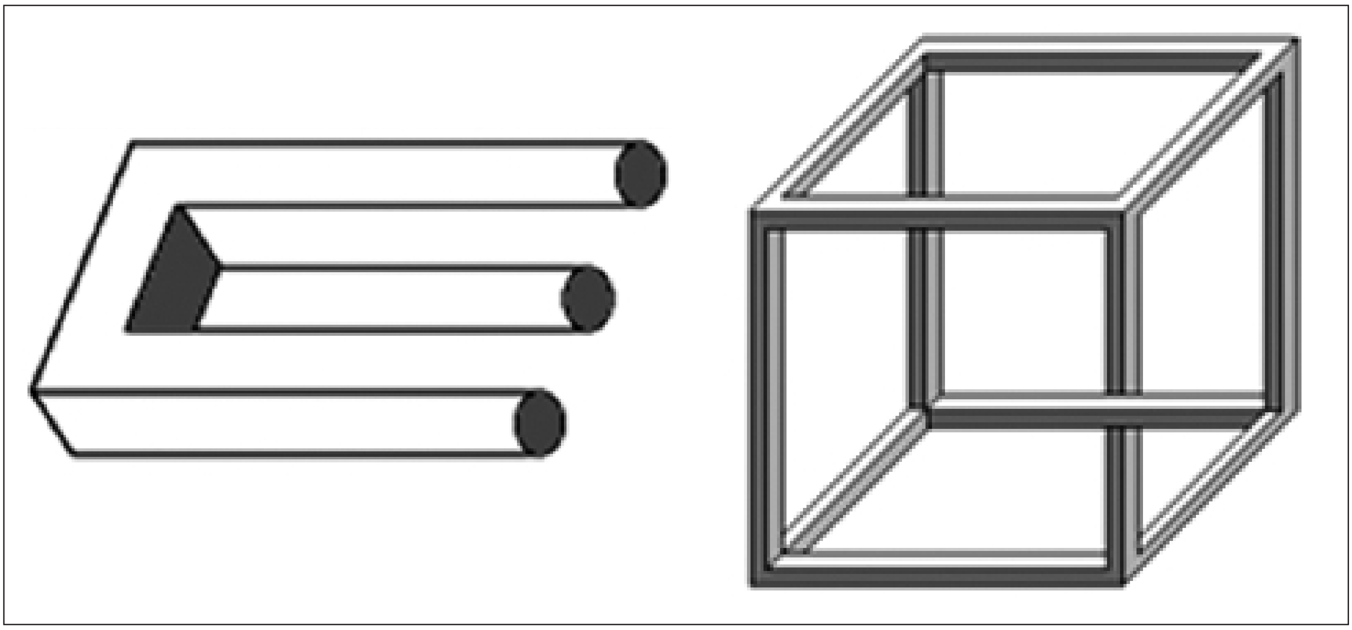


Figure 32. Drawings (after Escher) of a "three-pronged" fork and an impossible cube.

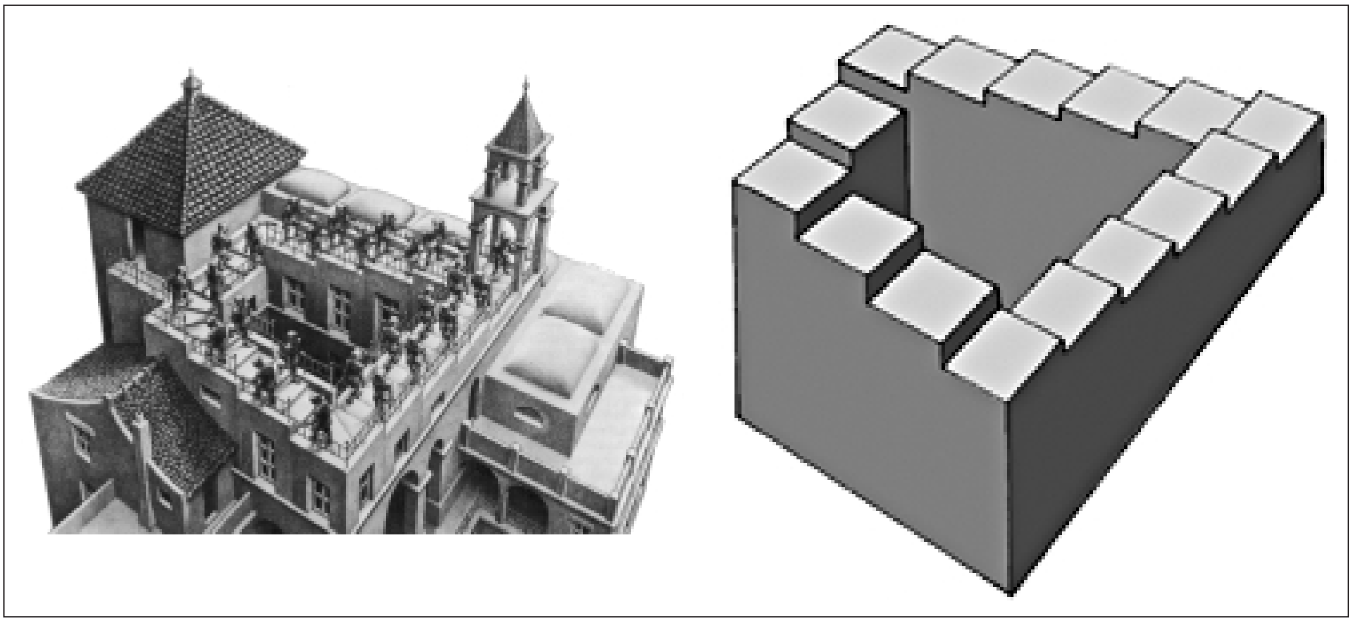
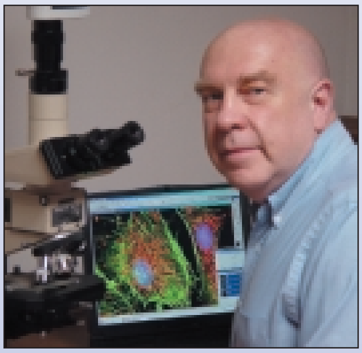


Figure 33. Escher's "climbing and descending" and a drawing of the endless stair.

ceived as being in front are grouped by angles that clearly place them at the back. In both of these cases, the local interpretation of the information is consistent, but no global resolution of the inconsistencies is possible.

Figure 33 shows another Escher drawing, called "climbing and descending." Locally the stairs have a direction that is everywhere clear and obvious. However, by clever use of perspective distortion, the steps form a closed path without top or bottom, so the path is endless and always in one direction. Several other geometric rivalries are also present in this drawing.

Article



John Russ is the author of The Image Processing Handbook, Computer Assisted Microscopy, Practical Stereology, Forensic Uses of Digital Imaging, Image Analysis of Food Structure, as well as many other books and papers. He has been involved in the use of a wide variety of microscopy techniques and the computerized analysis of microstructural images for nearly 50 years. One of the original founders of Edax International (manufacturer of X-ray analytical systems), and the past Research Director of Rank Taylor Hobson (manufacturer of precision metrology instruments), he has been since 1979 a professor in the Materials Science department at North Carolina State University. Now retired, he continues to write and lecture on topics related to image analysis.

"The first section of this three-part paper has emphasized the dependence of vision on local comparisons of brightness, color, orientation, and feature relationships. In the next part, comparisons over time are included to interpret motion, and comparisons over longer ranges are shown to influence judgments of distance. In addition, the tendency of people to see only a few things in a scene, and to see what they expect to see in a given context, is illustrated.

The third and concluding part deals with object recognition"

Seeing the Scientific Image, Part 2

John C. Russ

Materials Science and Engineering Department,
North Carolina State University, Raleigh, NC

It's about time

Comparison and inhibition operate temporally as well as spatially. The periphery of our vision is particularly well wired to detect motion. Comparison of the response at one point to that a short time previously is accomplished by a short time delay. Slightly more complicated connections detect motion of edges, with the ability to distinguish edges at different orientations. Gradual changes of brightness and motion, like gradual spatial variations, are ignored and very difficult to detect.

Temporal inhibition is not the same thing as adaptation or depletion. It takes a little while for our eyes to adapt to changes in scene brightness. Part of this is the response of the iris to open or close the pupil, letting more or less light into the eye. Part of the adaptation response is chemical, creating more amplification of dim light signals. The former requires many seconds and the latter many minutes to operate fully.

It is well established that staring at a fixed pattern or color target for a brief time will chemically deplete the rods or cones. Then looking at a blank page will produce an image of the negative or inverse of the original. Figure 34 shows a simple example. Stare fixedly at the center of the circle for about 60 seconds, and then look away. Because the color sensitive cones have been depleted, the afterimage of a circle composed of opposing colors will appear (green for red, yellow for blue, and so on).

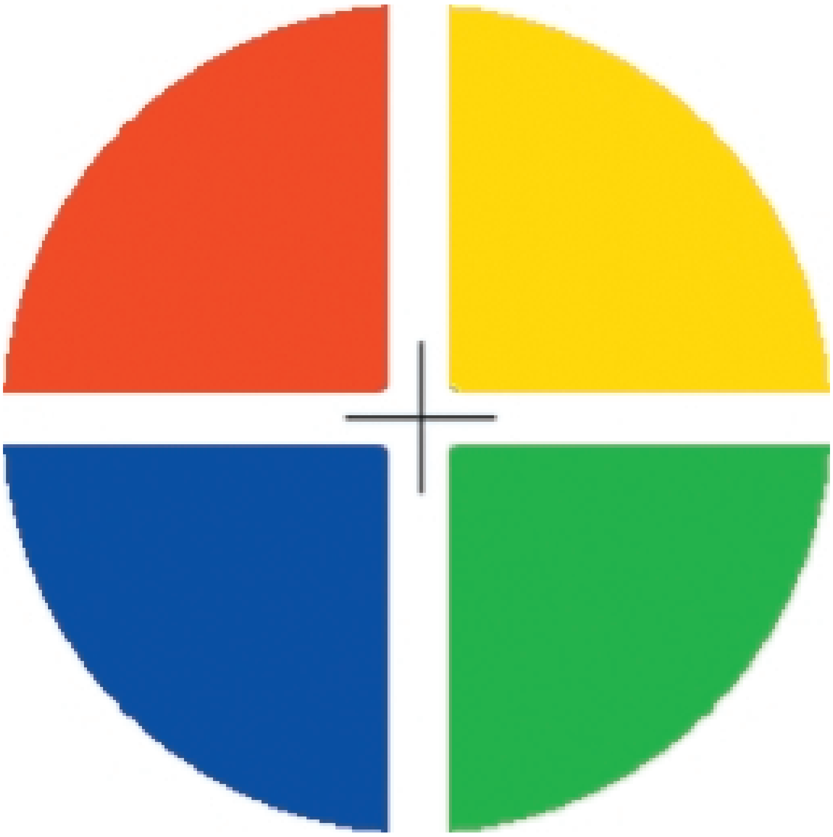


Figure 34. A target to demonstrate adaptation. Stare at the central cross for about a minute and then look away toward a blank sheet of paper. The complementary colors will be seen.

Motion sensing is obviously important. It alerts us to changes in our environment that may represent threats or opportunities. And the ability to extrapolate motion lies at the heart of tracking capabilities that enable us to perform actions such as catching a thrown ball. Tracking and extrapolation present a second-order opportunity to notice discontinuities of motion. If a moving feature suddenly changes speed or direction, that is also noticed. It is because of this ability to track motion and notice even subtle changes that the presentation of data as graphs is so useful, and why plots of the derivatives of raw data often reveal information visually.

Sequences of images are interpreted very differently if they occur at a sufficiently slow rate that they are seen as a sequence of individual pictures, or faster so that they present the illusion of continuous motion. Much of the blame for motion pictures and television can be assigned to Eadweard Muybridge, a photographer who was asked to settle a bet for Leland Stanford as to whether a galloping horse ever had all four feet off the ground. Muybridge set up a row of cameras with trip wires to photograph a horse as it galloped past, producing a sequence of pictures. Viewing them individually was enough to settle the bet, but Muybridge discovered that flipping through them rapidly gave the visual illusion of continuous motion. The movie industry began almost immediately. Figure 35 shows a sequence of twelve images of a trotting horse, taken by Muybridge.

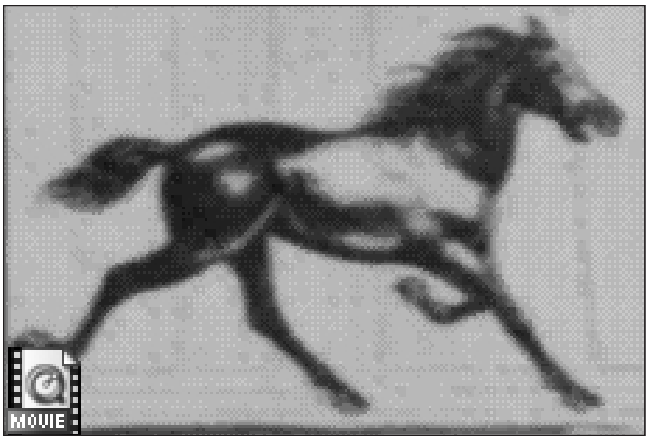


Figure 35. Muybridge's horse sequence
((this Quicktime movie can be downloaded from
http://DrJohnRuss.com/images/Seeing/Fig_35.mov)

In viewing a series of images, an important phenomenon called aliasing can occur. Generally we assume that a feature in one frame will correspond to a feature in the next if they are nearly the same in color and shape, and if they are close together. The familiar reversal of direction of the spokes on a turning wheel is an example of the visual error that results when a feature in one frame is matched to a different one in the next (Figure 36). From observing this phenomenon, it is a short step to stroboscopic imaging in which a series of pictures is taken at time intervals that match, or nearly match, the repetition rate of some phenomenon. This is commonly used to study turning or vibrating objects, falling water drops, and so on. Each image shows almost the same thing, albeit with a different tooth on the gear or a different water drop as the subject. We visually assume they are the same and perceive no change. By slightly varying the timing of the images (typically by controlling the rate at which the light source flashes) it is possible to extract subtle patterns and motions that occur within the faster repetitions.

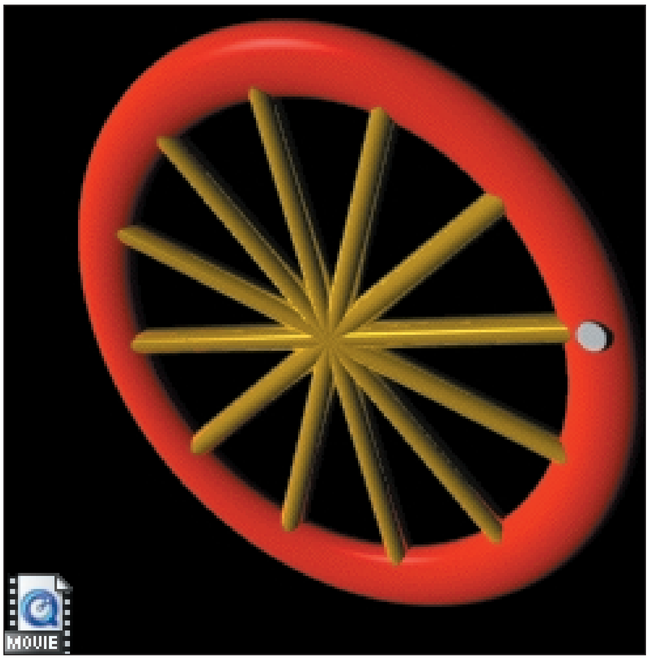


Figure 36. The wagon wheel effect
((this Quicktime movie can be downloaded from
http://DrJohnRuss.com/images/Seeing/Fig_36.mov)

If an image sequence is slower than about 15 frames per second, we don't perceive it as continuous. In fact, flickering images at 10-12 times a second can even induce epileptic fits in those with that affliction. But at higher rates, the temporal response of the eye and vision system sees continuous action. Movies are typically recorded at 24 frames per second, and television broadcasts either 25 (in Europe) or 30 (in the US) frames per second. In all of these cases, we interpret the sequential images as continuous. Incidentally, that is one of the problems with most video surveillance systems in use now. In order to record a full day's activity on a single VHS tape, the images are not stored at 30 frames per second but instead single frames (actually, single fields with only half the vertical resolution) are stored about every 4.5 seconds. The result is a series of still images from which some important clues are missing. We recognize people not only by still images but also by how they move, and the trajectories of motion are missing from the recording. We might detect posture, but not motion.

Just as human vision is best able to detect features or other causes of brightness or color variation over a

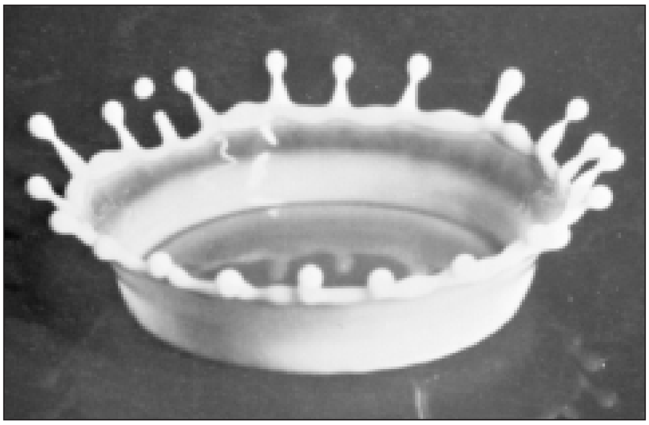


Figure 37. Doc Edgerton's high speed photograph of a milk drop creating a splash.

relatively narrow range of sizes, so it deals best with events that occur over a relatively narrow range of times. Very short duration events are invisible to our eyes without devices such as high speed photography to capture them (Figure 37). Both high speed imaging and stroboscopic imaging techniques were pioneered by Doc Edgerton, at MIT.

Likewise, events that take a long time (minutes) to happen are not easy to examine visually to detect changes. Even side-by-side images of the clock face in Figure 38 don't reveal all ot the differences, and when one of the images must e recalled from memory the results are even poorer. Capturing the images in a computer and calculating the difference shows clearly the motion of the minute hand and even the very slight shift of the hour hand.

Considering the sensitivity of vision to motion, it is astonishing that our eyes are actually in nearly constant motion. It is only in the Fovea that high resolution viewing is possible, and we rotate our eyes in their sockets so that this small high resolution region can view many individual locations in a scene, one after another. Flicking through the points in a scene that seem "interesting" gathers the information that our minds require for interpretation and judgment. Most of the scene is never actually examined, unless the presence of edges, lines, colors, or abrupt changes in color, brightness, texture or orientation make locations interesting. Somehow, as our eyes move and different images fall onto the retina every few hundred milliseconds, our minds sort it all out and plug the information into the perceptual scene that is constructed in our head. Although the entire image shifts on the retina, it is only relative motion within the scene that is noticed. This motion of the eye also serves to fill in the blind spot, the location on the retina where the connection of the optic nerve occurs, and where there are no light sensors. Experiments that slowly move a light through the visual field while the eyes remain fixed on a single point easily demonstrate that this blind spot exists, but it is never noticed in real viewing because eye motion picks up enough information to fill the perceptual model of the scene that we actually interpret. But there may be a great deal of information in the actual scene that we do not notice. Clearly there is a lot of interpretation going on.

The perceived image of a scene has very little in common with a photographic recording of the same scene. The latter exists as a permanent record and each part of it can be examined in detail at a later time. The mental image of the scene is transitory, and much of it was filled with low resolution information from our visual periphery, or was simply assumed from memory of other similar scenes. Scientists need to record images rather than just view them, because it is often not obvious until much later which are the really important features present (or absent).

The evolution of the structure and function of the eye and the brain connections that process the raw data have been a response to environment and challenges of survival. Different animals clearly have different needs and this has resulted in different types of eyes, as noted above, and also different kinds of processing. Apparently the very few but extremely specific bits of information that the fly's eye sends to the fly's brain are enough to trigger appropriate and successful responses (for example, a looming surface triggers a landing reflex in which the fly re-orient's its body so the legs touch first). But for most of the "higher" animals the types of information gathered by the visual system and the interpretations that are automatically applied are much more diverse.

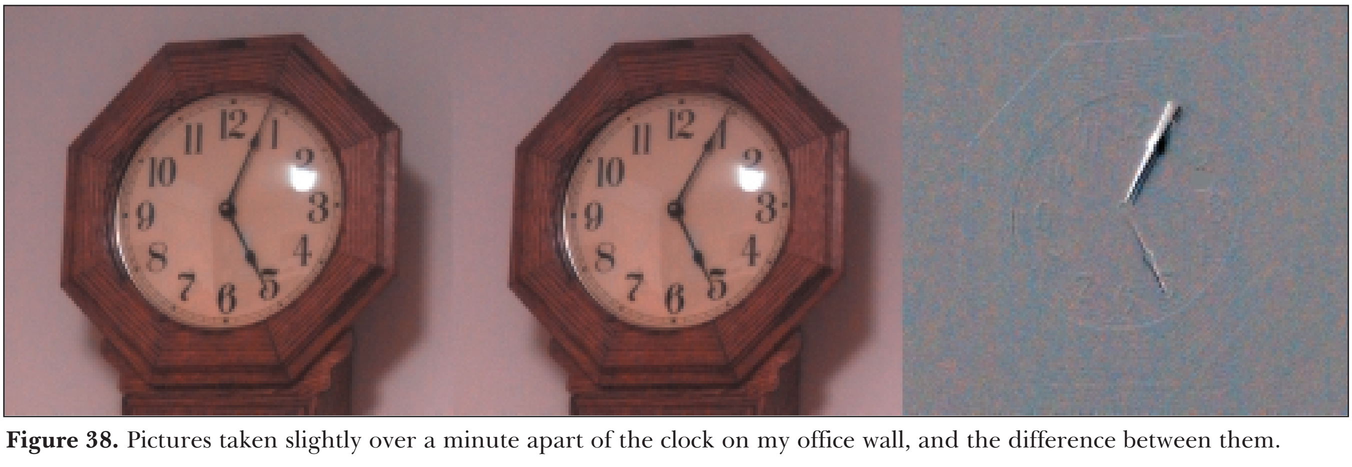


Figure 38. Pictures taken slightly over a minute apart of the clock on my office wall, and the difference between them.

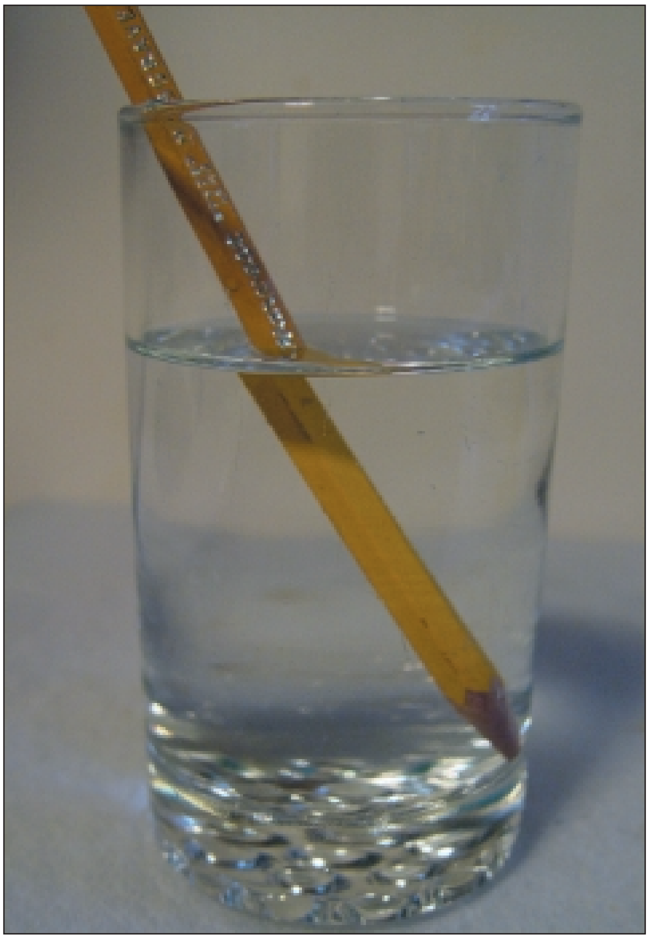


Figure 39. The apparent bend in the pencil and its magnification are familiar optical effects.

For example, when people see a pencil placed in a glass of water, the pencil appears to be bent at an angle (Figure 39). Intellectually we know this is due to the difference in the refractive indices of water and air, but the view presented to our brain still has the bend. South Pacific islanders who have spent their lives fishing with spears in shallow water learn how to compensate for the bend but the image that their eyes present to their brains still includes the optical effect. There is evidence that fishing birds like the heron have a correction for this offset built in to their visual system and that from the point of view of the perceived image they strike directly toward the fish in order to catch it. Conversely, there are species of fish that see bugs above the water and spit water at them to knock them down and catch them for food. This requires a similar correction for the optics.

It may seem strange that computation in the brain could distort an image in exactly the proper way to correct for a purely optical effect such as the index of refraction of light in water. But this is different only in degree from the adjustments that we have already described, in which human vision corrects for shading of surfaces and illumination that can vary in intensity and color. The process combines the raw visual data with a lot of "what we know about the world" in order to create a coherent and usually accurate, or at least accurate enough, depiction of the scene.

Humans are very good at tracking moving objects and predicting their path taking into account air resistance and gravity (for instance, an outfielder catching a fly ball). Most animals do not have this ability but a few have learned it. My dog chases a thrown ball by always running toward it's present position, which produces a path that is mathematically a tractrix. It works, it doesn't take much computation, but it isn't optimum. My cat, on the other hand, runs in a straight line toward where the ball is going to be. She has solved the math of a parabola, but having pounced on the ball, she won't bring it back to me, as the dog will.

What we typically describe in humans as "eye-hand" coordination involves an enormous amount of subtle computation about what is happening in the visual scene. A professional baseball or tennis player who can track a ball moving at over 100 miles per hour well enough to connect with it and hit it in a controlled fashion has great reflexes, but it starts with great visual acuity and processing. It was mentioned earlier that the human eye has some 150+ million light sensors, and for each of them some 25-50,000 processing neurons are at work extracting lots of information that evolution has decided we can use to better survive.

The third dimension

Most people have two functioning eyes, and have at least a rudimentary idea that somehow two eyes allow stereoscopic vision that provides information on the distance of objects that we see. Lots of animals have eyes located on the sides of their heads where the overlap in the two fields of view is minimal, and don't rely on stereo vision very much. It is mostly predators and particularly animals that live by leaping about in trees that have their eyes forward facing, indicating that stereoscopic vision is important. But it is by no means the only way by which distances are determined, and it isn't a particularly quantitative tool.

Humans use stereo vision by rotating the eyes in their sockets to bring the same feature to the fovea in each eye (as judged by matching that takes place in the visual cortex). It is the feedback from the muscles to the brain that tell us whether one feature is closer than another, depending on whether the eyes had to rotate in or out as we directed our attention from the first feature to the second. Notice that this is not a measurement of how much farther or closer the feature is, and that is works only for comparing one point to another. It is only by glancing around the scene and building up a large number of two-point comparisons that our brain constructs a map of the relative distances of many locations.

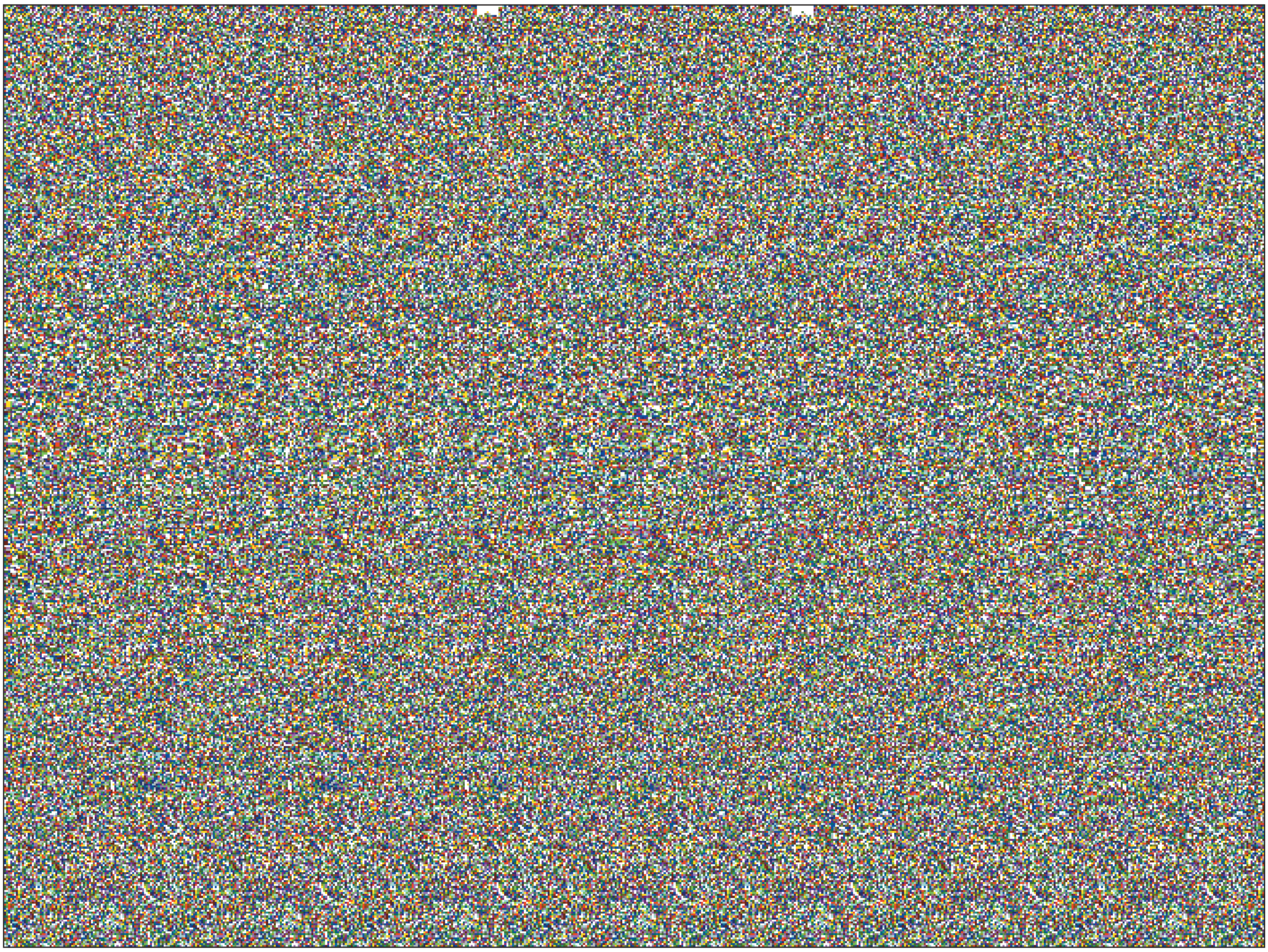


Figure 40. Random dot stereogram showing a torus on a plane (stare and the image and the eyes will pick out the matching patterns and fuse them into an image).

Stereoscopy can become confused if there are several similar features in the image, so that multiple matches (corresponding to different apparent distances). This happens rarely in natural scenes, but can create problems in microscopy of repetitive structures.

One of the fascinating discoveries about stereopsis, the ability to fuse stereo pair images by matching details from the left and right eye views, is the random-dot stereogram (Figure 40). Bela Julesz showed that the visual system was able to match patterns of dots that to a single eye appeared chaotic and without structure, to form stereo images. Slight lateral displacements of the dots are interpreted as parallax and produce depth information.

Sequential images produced by moving a single eye can also produce stereoscopic depth information. The relative sideways motion of features if we shift our head from side to side is proportional to distance. One theory holds that snakes, whose eyes are not well positioned for stereo vision, move their heads from side to side to better triangulate the distance to strike.

Stereoscopy only works for things that are fairly close. At distances beyond about 100 feet, the angular differences become too small to notice. Furthermore, there are plenty of people who, for one reason or another, do not have stereo vision (for example, this is a typical consequence of childhood amblyopia, or lazy eye), who still function quite well in a three-dimensional world, drive cars, play golf, and so on. There are several other cues in images that are used to judge distance.

If one object obscures part of another, it must be closer to our eye. Precedence seems like an absolute way to decide the distance order of objects, at least those that lie along the same line of sight. But there built-in assumptions of recognition and simple shape of objects, as shown in the example in Figure 41, whose violations create an incorrect interpretation.

Relative size also plays an important role in judging distance. Many of the things we recognize in familiar scenes have sizes - most typically heights - that fall into narrow ranges. Closely related is our understanding of the rules of perspective - parallel lines appear to converge as they recede (Figure 42). In fact, the presence of such lines is interpreted visually as corresponding to a surface that extends away from the viewer, and irregularities in the straightness of the

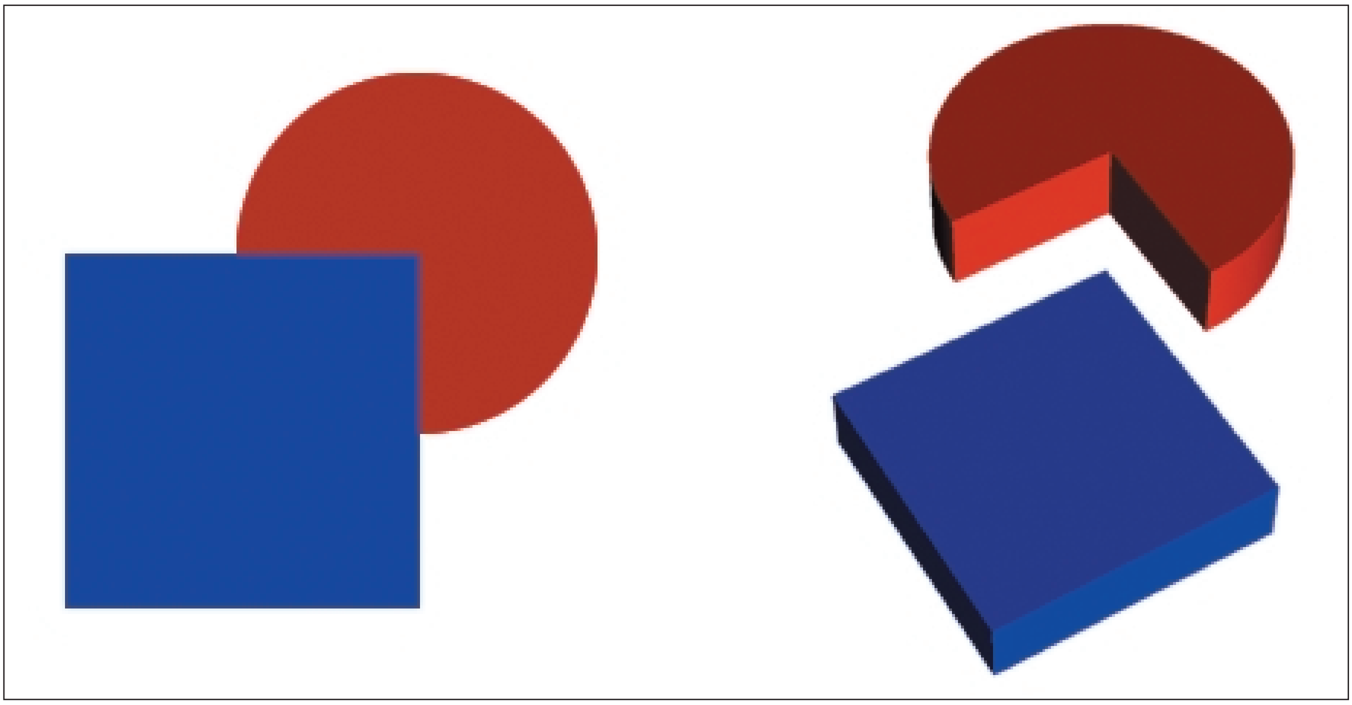


Figure 41. Left: Obviously the blue square is in front of the red circle. Right: But it may not be a circle, and viewed from another angle we see that the red feature is actually in front of the blue one.

lines are interpreted as representing bumps or dips on the perceived surface. Driving through the countryside looking at plowed fields provides a simple example.

By comparing the apparent size of features in our visual field we judge the relative distance of the objects and of things that appear to be close to them. But again, the underlying assumptions are vital to success, and violations of the straightness of alignments of features or the constancy of sizes produces illusory interpretations (Figure 43).

A simple extension of the assumption of size constancy for major features uses the sizes of marks or features on surfaces to estimate distance and angle. It is logical, and indeed often correct, to assume that the marks or texture present on a surface are random and isotropic, and that visual changes in apparent size or aspect ratio indicate differences in distance or orientation (Figure 44). Once again, violations of the underlying assumptions lead us to the wrong conclusions.

There are other clues that may be present in real scenes, although they are less relevant to the viewing of images from microscopes or in the laboratory. For instance, atmospheric haze makes distant features appear more blue (or brown, if the haze is smog) and less sharp. Renaissance painters, who mastered all of these clues, represented atmospheric haze in scenes along with cor-

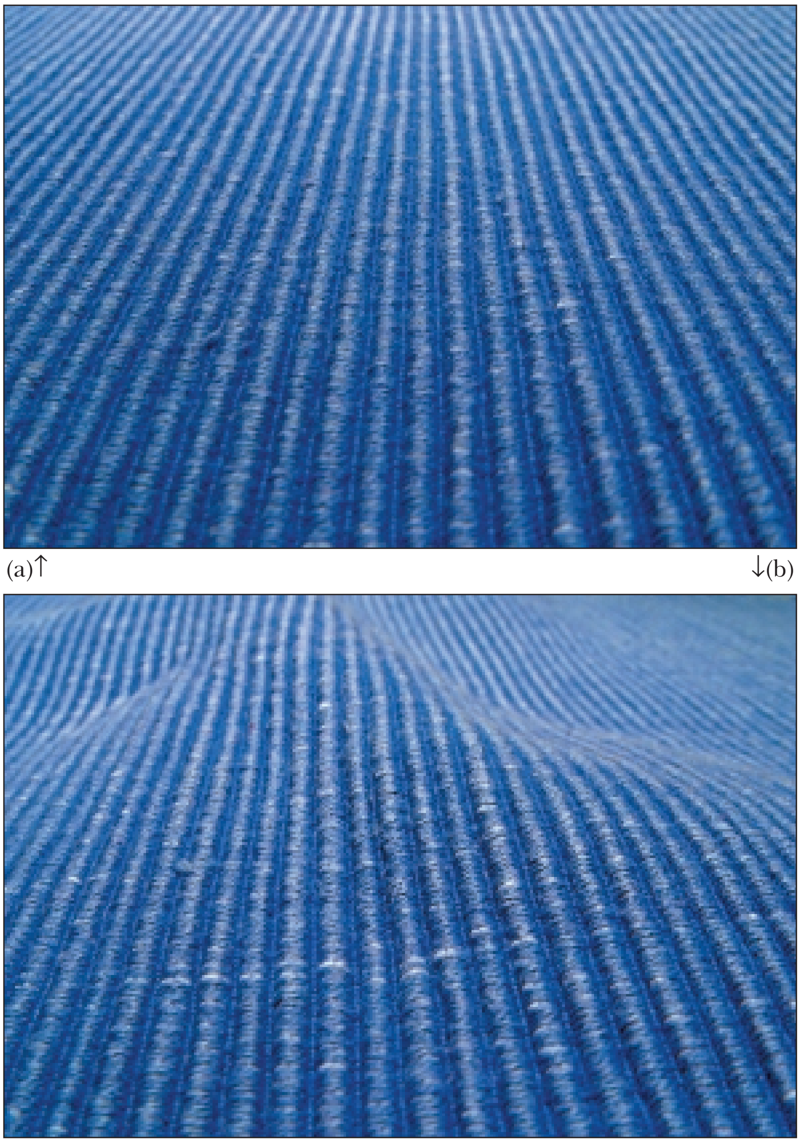


Figure 42. Converging lines are interpreted as parallel lines that converge according to the rules of perspective, and so the surface is perceived as receding from the viewer. Straight lines imply a flat surface (a), while irregularities are interpreted as bumps or dips in the perceived surface (b).

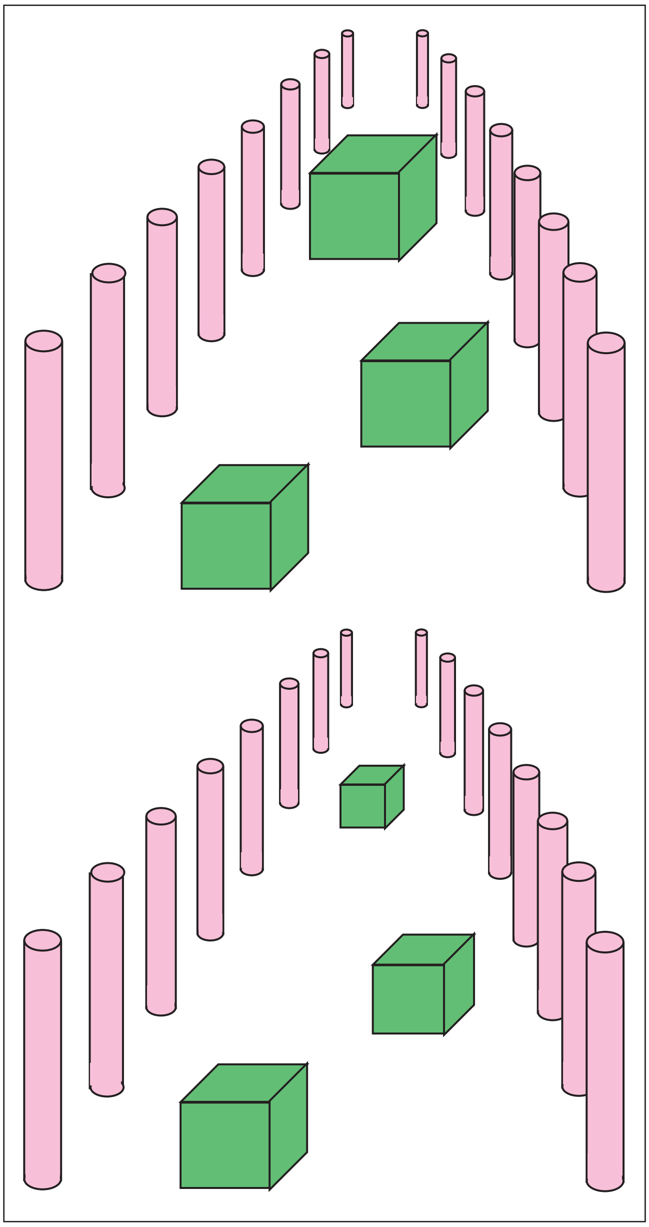


Figure 43. In these illustrations, the expectation of distance is established by assuming the pink posts are constant in size and arranged in straight parallel lines, whose apparent convergence is a function of perspective. In the bottom illustration, the appearance of the green boxes is consistent with this interpretation. In the top illustration it is not, and we must either conclude that the boxes differ in size or the posts do not conform to our expectation.

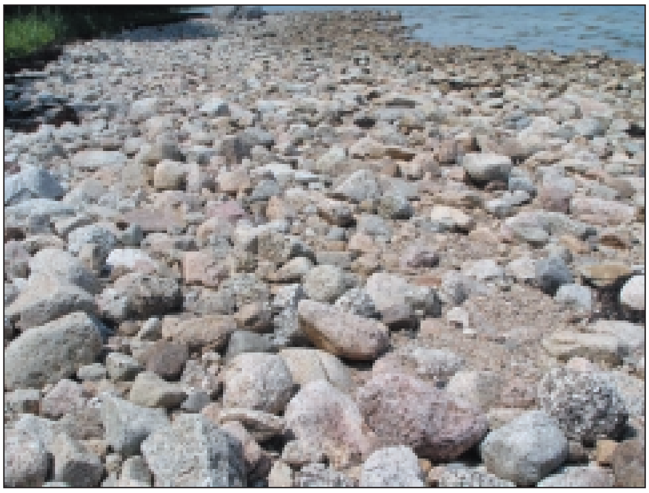


Figure 44. Rocks on a beach. Assuming that the rocks are similar in size and round on the average informs us of the viewing angle and the distance to farther locations.

rect geometric perspective. Working from a known geometry to a realistic representation is a very different task than trying to extract geometric information from a scene whose components are only partially known, however.

How versus What

Several very plausible models have been put forward for the algorithms functioning in the eye and other portions of the visual pathway. The first few layers of neurons in the retina are connected in ways that can account for mechanisms behind local and temporal inhibition, and the interleaving of information from right and left eyes in the visual cortex is consistent with the fusion of two images for stereopsis. Such models serve several purposes - they can be tested by physiological probes and external stimuli, and they form the basis for computer techniques that attempt to extract the same information from images. In fact, they serve the latter purpose even if they turn out to be failures as actual descriptions of the functioning of the neurons. But while they may be effective at describing HOW at least some parts of the visual system work, because they work at the level of bits and pieces of the image and not the Gestalt or information level, they don’t tell us much about WHAT we see.

Several years ago I was retained as an expert witness in a major criminal trial. The issue at hand was whether a surveillance video tape from the scene of a murder was useful for the identification of the suspects. The images from the tape had been extensively computer-enhanced and were shown to the jury, who were invited to conclude that the current appearance of the defendants couldn’t be distinguished from those images. In fact, the images were so poor in both spatial and tonal resolution that they couldn’t be distinguished from a significant percentage of the population of the city in which the murders took place, and it was the job of the defense to remind the jury that the proper question was whether the pictures contained enough matching information to identify the defendants “beyond a reasonable doubt.” It was very interesting in this case that none of the numerous witnesses to the crime were able to pick any of the defendants out from a lineup. The human eye is indisputably a higher resolution, more sensitive imaging device than a cheap black and white surveillance video camera. But for a variety of reasons the humans present could not identify the perpetrators.

In that trial I was accepted by the court as an expert both in the field of computer-based image processing (to comment on the procedures that had been applied to the camera images) and on human perception (to comment on what the people present might have been able to see, or not). The point was raised that my degrees and background are not in the field of physiology, but rather physics and engineering. How could I comment as an expert on the processes of human vision? The point was made (and accepted by the court) that it was not an issue of how the human visual system worked, at the level of rhodopsin or neurons, that mattered, but rather of what information human vision is capable of extracting from a scene. I do understand what can be seen in images, because I've spent a lifetime trying to find ways for computers to extract some of the same information (using what are almost certainly very different algorithms). Accomplishing that goal, by the way, will probably require a few more lifetimes of effort.

In fact, there has often been confusion over the difference between the How and the What of human vision, often further complicated by questions of Why. In describing a computer algorithm for some aspect of image analysis, the explanation of the steps by which information is extracted (the How) is intimately bound up in the computed result (the What). But in fact, the algorithm may not be (in fact usually is not) the only way that information can be obtained. Many of the important steps in image analysis have several more-or-less equivalent ways of extracting the same result, and moreover each of them can typically be programmed in quite a few different ways to take advantage of the peculiarities of different computer architectures. And, of course, no one claims that any of those implementations is identical to the processing carried out by the neurons in the brain.

David Marr, in his final book "Vision" (Freeman, 1982) has pointed out very forcefully and eloquently that confusing the How and the What had led many researchers, including himself, into some swampy terrain and dead ends in the quest for an understanding of vision (both human and animal). Mapping the tiny electrical signals in various parts of the visual cortex as a function of stimuli presented to the eye, or measuring the spectral response of individual rod or cone cells, is certainly an important part of eventually understanding the How of the visual system. And it is an experiment that is performed at least in part because it can be done, but it isn't clear that it tells us very much about the What. On the other hand, tests that measure the response of the frog's eye to a small dark moving target invite speculation about the Why (to detect a moving insect - food).

Researchers have performed many experiments to determine what people see, usually involving the presentation of artificial stimuli in a carefully controlled setting and comparing the responses as small changes are introduced. This has produced some useful and interesting results, but falls short of addressing the problem of visual interpretation of scenes. The key is not just that people can detect ("see") a certain stimulus, but that they can interpret its meaning in a complex scene. It might be better to substitute the word "interpret" for "see" to emphasize that the individual cues in images are only important for understanding the world when they are combined and processed to become a semantic representation. In other words, we do have to turn that picture into its "thousand word" equivalent. For that purpose, it is often more revealing to study the errors that humans make in looking at whole images. This includes, but is not limited to, various kinds of visual illusions.

There are also important clues in what artists have portrayed (or left out) of representational paintings and even cartoons. By exaggerating a few selected features into a caricature, for example, editorial cartoonists create very distorted but extremely recognizable representations of familiar political figures. For many people, such cartoons may represent a greater truth about the person than an actual photograph (Figure 45).

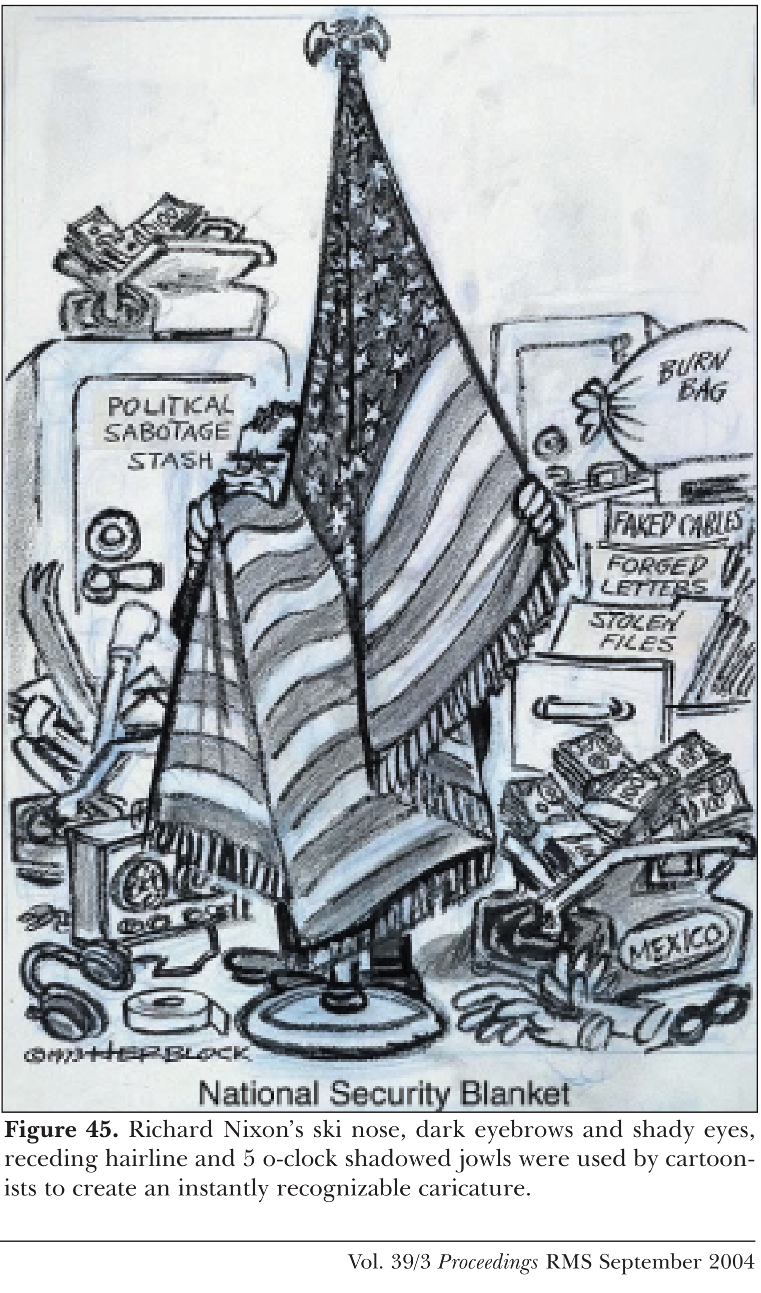


Figure 45. Richard Nixon's ski nose, dark eyebrows and shady eyes, receding hairline and 5 o-clock shadowed jowls were used by cartoonists to create an instantly recognizable caricature.

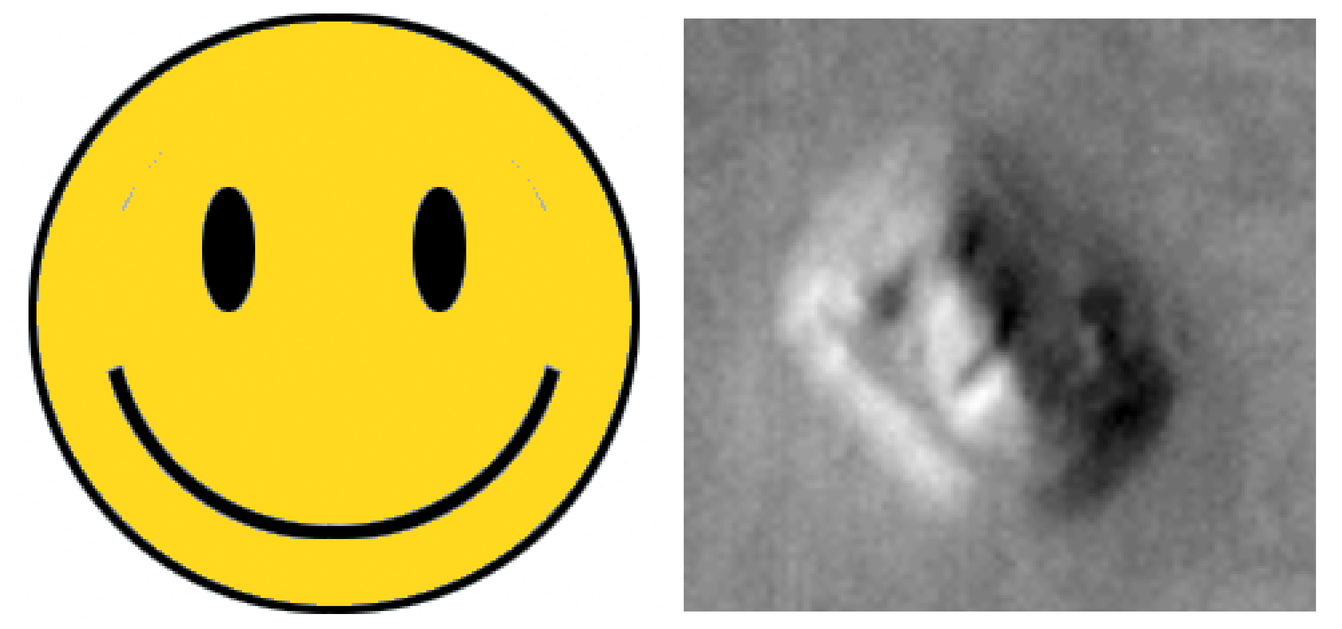


Figure 46. It takes very few cues to trigger the recognition of a face: a) the ubiquitous happy face; b) the "face on Mars" which appears only if the viewing angle and lighting are correct.

One problem that plagues eyewitness testimony and identification is that we tend to see (i.e., pick out from a scene) things that are familiar (i.e., already have mental labels). One facility that is hard-wired into our brains, just like the ability of the frog to spot a bug, is finding faces. Babies find and track faces from birth. We are so good at it that even with just a few clues, like two eyes and a mouth, we see a face, whether it is real or not. The ubiquitous "smiley face" cartoon has enough information to be recognized as a face. So does a mountain on Mars, when illuminated and viewed a certain way (Figure 46).

But to recognize a particular face, for instance as grandmother, we need a lot more clues. Computer programs that perform facial recognition use ratios of dimensions, for instance the ratio of the distance between the eyes to that between the tips of the ears, or the distance between the mouth and chin to the distance from the tip of the nose to the chin. The advantage of ratios is that they are insensitive to the size of the image, and to a considerable extent to orientation or point of view. But that is an algorithm and so addresses the How rather than the What. It seems likely that human facial recognition uses more or different clues, but certainly altering those proportions by even a few percent changes a face so that it becomes unrecognizable (Figure 47).

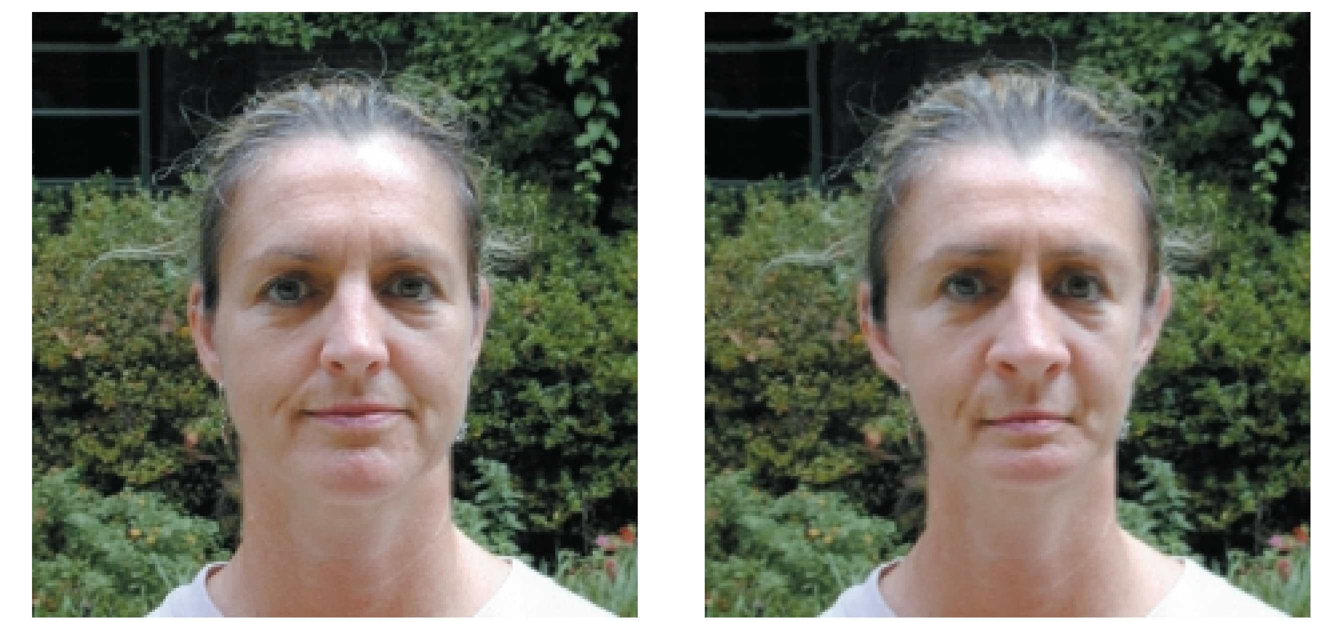


Figure 47. Altering the ratios of dimensions (such as the horizontal distance between the eyes, ears, width of mouth, etc., or vertical dimensions such as length of nose, distance from mouth to chin, height of forehead, etc.) strongly affects our ability to recognize faces.

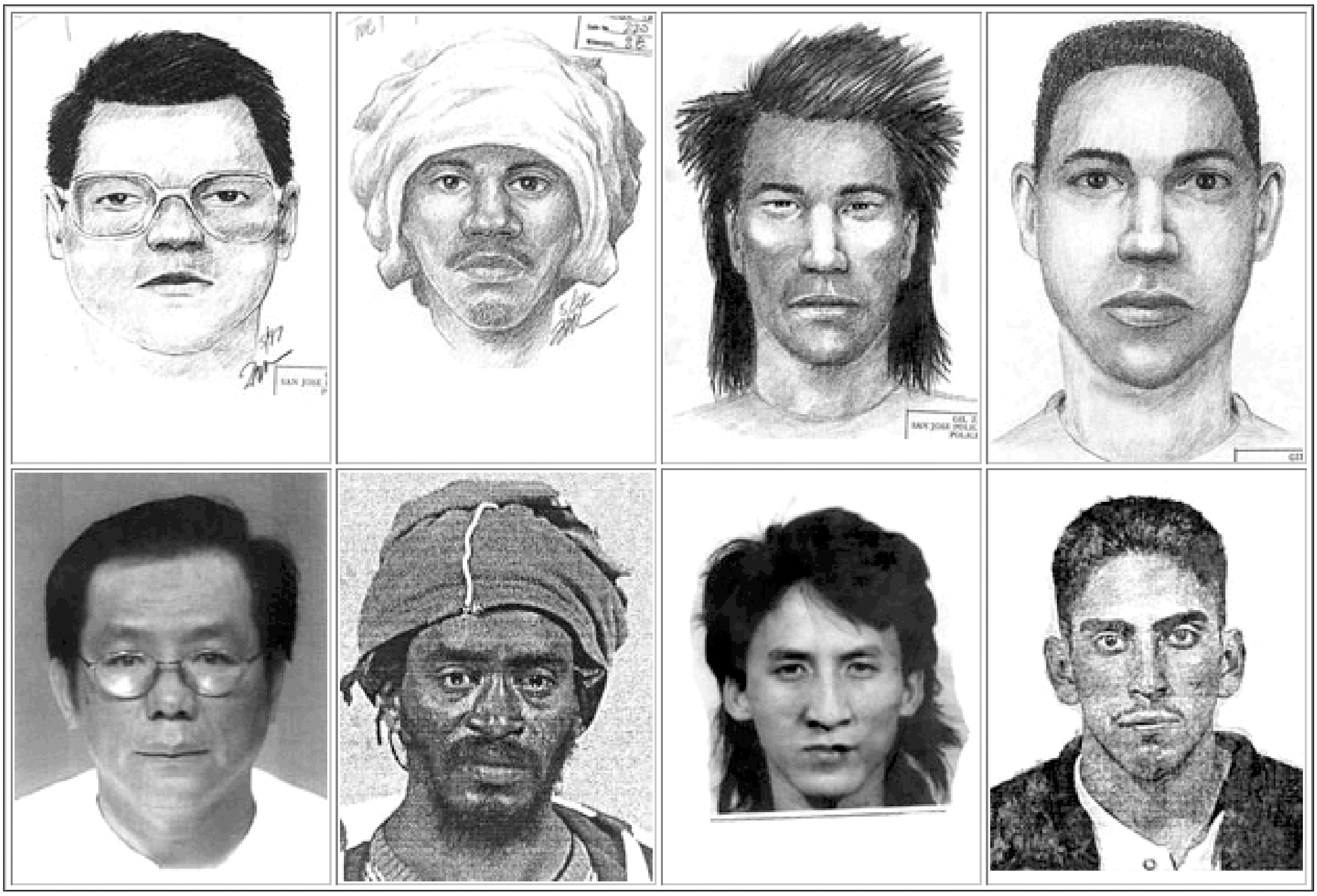


Figure 48. Examples of police artist sketches and photos of the actual persons.

Police artists routinely produce sketches from eyewitness descriptions. Comparing these pictures to actual photographs of perpetrators after capture suggests that only a few characteristics of the face are likely to be noticed, and turned into a mental caricature rather than an actual representation (see the examples in Figure 48). And differences in race between witness and perpetrator make it especially difficult to pick out those characteristics that are likely to identify the person. We learn to pick out the particular kinds of details that are most useful in identifying those familiar to us, and these are not very helpful for other races ("they all look alike").

Finally, when something or someone is recognized (rightly or wrongly) in an image, our minds mentally endow the semantic representation of that person or object with the full set of characteristics that we remember from past encounters. A typical example would be seeing a friend's face from the side, but "knowing" that there is a mole on the other cheek and believing we had seen that this time as well. That leads to a lot of eyewitness problems. If a witness thinks they have recognized someone or something, they will often testify with confidence and honesty that they have seen things that were actually not present. This can include even highly specific items like articles of clothing, tattoos, etc. One witness was sure that a car in a hit and run case had a particular bumper sticker on the back, when in fact she had not been in a position to see the back of the car, because she was familiar with a car of the same model and color that did have such a bumper sticker.

We all do this, unconsciously. When you pass a neighbor's house, if you glimpse someone mowing the lawn and you "expect" it to be the teenage son, you are likely to "see" details of his appearance, haircut, clothing, etc., that may not be visible, or may not even be there - it might not even be the right person. Usually this process is helpful because it gives us a sense of place and situation that is most often correct without requiring additional time or effort. A few mistakes will be made, but fortunately they aren't usually serious ones and the consequences are rarely more than momentary embarrassment. Many decades ago when my eldest son was a preschooler, I shaved off a mustache that I had worn since before his birth. He did not notice for two days, until it was pointed out to him (and then he was upset at the change).

Seeing what we "know" is present, or at least expect to be present, is common. A colleague of mine, who for years helped in teaching courses on image analysis, has a favorite picture of herself holding a dearly loved pet, now deceased. Unfortunately the dog is black, she is wearing a black sweater, and the photographic print is very dark (and the negative, which would have a greater dynamic range, is not available). She has challenged us for years to process that image to show the dog that she sees when she looks at the picture, but we've failed because there is just nothing there in terms of the pixel values - they are all the same shade of near black. One of my students took a copy of the scanned print and painstakingly drew in a plausible outline of the correct breed of dog. Her immediate response was "That's not the right dog!" She has a stored image in her mind that contains information not available to anyone else who looks at the image, and she believes she can see that information when she looks at the picture. Certainly her mind sees it, if not her eyes.

The extension of this process to scientific image analysis is obvious and should be of great concern. We see what we expect to see (things for which we have existing mental labels), fail to see or recognize things that are unfamiliar, misjudge things for which we do not have an appropriate set of stored clues, and truly believe that we have seen characteristics in one image that have been seen in other instances that are remembered as being similar. That's what it means to be human, and those are the tendencies that a careful scientific observer must combat in analyzing images.

Image compression

There are some important lessons about human vision to be found in the rapid acceptance of digital still and video cameras. All consumer cameras and many higher end cameras store images in a compressed format because memory is expensive and also smaller files can be saved more quickly (more than enough improvement to make up for the time needed to carry out the compression). People seem willing to pay for high resolution multi-megapixel cameras and then try to compress the image by a factor of 10, 20 or more to produce a file that can be transmitted efficiently over the internet.

Compression techniques such as MPEG for video and JPEG for still pictures are widely used and little questioned. In addition to MPEG (Moving Pictures Expert Group) compression, a variety of codecs (compressor-decompressor) are available for Apple's Quicktime and Macromedia's Flash software. The original JPEG (Joint Photographers Expert Group) technique using a discrete cosine transform has been joined by wavelet and fractal methods.

All of these methods achieve compression by leaving out some of the information in the original image; technically they are "lossy" compression techniques. The intent of the compression is to preserve enough information to enable people to recognize familiar objects. Most of the techniques depend to some extent on the characteristics of human vision to decide what should be kept and what can be modified or left out. Some, like fractal compression, replace the actual details with other detail "borrowed" from elsewhere in the image, on the theory that any fine detail will fool the eye.

Compression discards what people don't easily see in images. Human vision is sensitive to abrupt local changes in brightness, which correspond to edges. These are kept although they may shift slightly in location, and in magnitude. On the other hand, absolute brightness is not visually perceived so it is not preserved. Since changes in brightness of less than a few percent are practically invisible, and even larger variations cannot be seen if they occur gradually over a distance in the image, compression can eliminate such details.

Color information is reduced in resolution because boundaries are primarily defined by changes in brightness. The first step in most compression schemes is to reduce the amount of color information, either by averaging it over several neighboring pixels or by reducing the number of colors used in the image, or both. Furthermore, color perception is not the same in all parts of the visible spectrum. We cannot discriminate small changes in the green range as well as we can other colors. Also, gradual changes in color, like those in brightness, are largely invisible, only sharp steps are noticed. So the reduction of color values in the image can be quite significant without being noticeable.

It is also possible to reduce the size of video or movie files by finding regions in the image that do not change, or do not change rapidly or very much. In some cases, the background behind a moving object can be simplified, even blurred, while the foreground feature can be compressed because we don't expect to see fine detail on a moving object. Prediction of the locations in an image that will attract the eye (sometimes called "interesting points" and usually associated with high local contrast, or familiar subjects - where do your eyes linger when looking at a picture of a movie star? what parts of the image don't you notice?) and cause it to linger in just a few areas allows other areas to be even further compressed.

Certainly it can be argued that this type of compression works below the threshold of visual discrimination most of the time, and does not prevent people from recognizing familiar objects. But that is exactly the problem: compression works because enough information remains to apply labels to features in the image, and those labels in turn cause our memories to supply the details that are no longer present in the picture. The reason for recording images in scientific studies is not to keep remembrances of familiar objects and scenes, but to record the unfamiliar. If it is not possible to know beforehand what details may turn out to be important, it is not wise to discard them. And if measurement of features is contemplated (to measure size, shape, position or color information), then lossy compression, which alters all of those values, must be avoided.

It is not the point of this section to just make the rather obvious case that compression of digital images is extremely unwise and should be avoided in scientific imagery. Rather, it is to shed illumination on the fact that compression is only acceptable for snapshots because human vision does not notice or depend upon very much of the actual contents of an image. Recognition requires only a few clues, and ignores much of the fine detail.

A world of light

Our eyes are only one part of the overall system involved in producing the sensory input to the visual cortex. It is easy to overlook the importance of the light source and its color, location, and brightness. A few concerns are obvious. When shopping for clothes, furniture, or other items, it is best not to rely on their appearance under the artificial lighting in the store, but to see how the colors appear in sunlight. The difference in color temperature of the light source (sunlight, incandescent lighting, fluorescent lighting) can produce enormous differences in the visual appearance of colors. And don't even think about trying to guess at colors using the illumination from a sodium street light, which is essentially monochromatic and provides no clues to color at all.

In a laboratory setting, such as the use of a light microscope, color judgments can be similarly affected by small variations in the color temperature of the bulb, which depends very sensitively on the voltage applied to it (and also tends to change significantly over the first few and last few hours of use as the filament and its surface undergo physical alterations). Simply reducing the illumination (e.g., to take a photo) by turning down the voltage will change the colors in the image. This happens whether we are imaging the light that is transmitted (not reflected or absorbed) through a thin sample, as in the transmission light microscope, or the light reflected from a surface, as in macroscopic imaging. But generally it is the latter case that our brains are prepared to interpret.

For real world scenes, the light source may be a single point (e.g., sunlight or a single bulb), or there may be multiple sources. the source may be highly localized or it may be extended. Lighting may be direct or indirect, meaning that it may have been reflected or scattered from other surfaces between the time it leaves the source and reaches the object. All of these variables affect the way the object will appear in the final image.

The surface of the object also matters, of course. Most of the light that is not transmitted through or absorbed within the object is scattered from an extremely thin layer just at the surface of the object. For a perfect metal, this happens exactly at the surface, but most materials allow the light to penetrate at least a short distance beneath the surface. It is the variation of absorption within this thin layer for different light wavelengths, and the variation of penetration with wavelength, that gives an object color. For instance, preferential absorption of green light will cause an object to appear purple. Ideal metals, for which there is no light penetration, have no color (the colors of gold, copper and silver result from a very complex electronic structure that actually allows slight penetration).

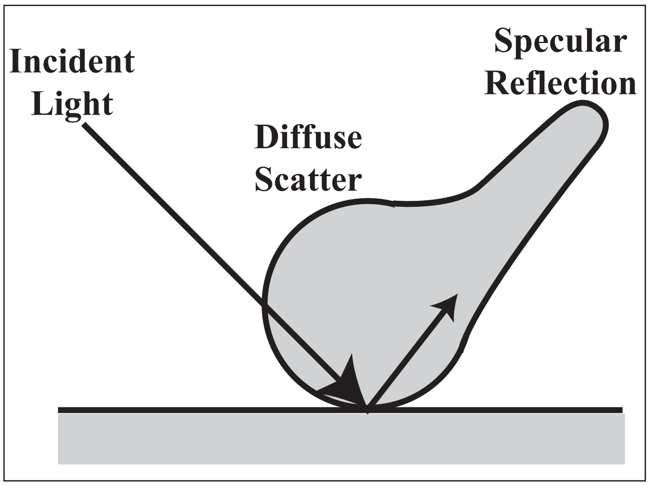


Figure 49. The relative amount of diffuse scattering and specular reflection, the angular breadth of the specular reflection, and the total amount of light that is scattered rather then transmitted or absorbed, all of which may vary with wavelength, determine the appearance of surfaces.

The fraction of the incident light that is reflected or scattered is measured by the surface albedo. A very dark object may absorb as much as 90% of the incident light, whereas a very bright one may absorb only a few percent. The interaction of the light with the object typically includes a mixture of diffuse and specular reflection. The diffuse component sends light in all directions, more of less following a cosine pattern as shown in Figure 49. The specular component sends light in the particular direction of mirror reflection with an angle to the local surface normal equal to the incident angle. The specularity of the surface is defined by the fraction of the light that reflects at the mirror angle and the narrowness of the reflected beam.

Computer programs that generate rendered surface images from measurements and shape information use models that correspond to the behavior of typical materials. As shown in the Figure 50, it is possible to change the appearance of the surface and our judgment of its composition by altering the specularity.

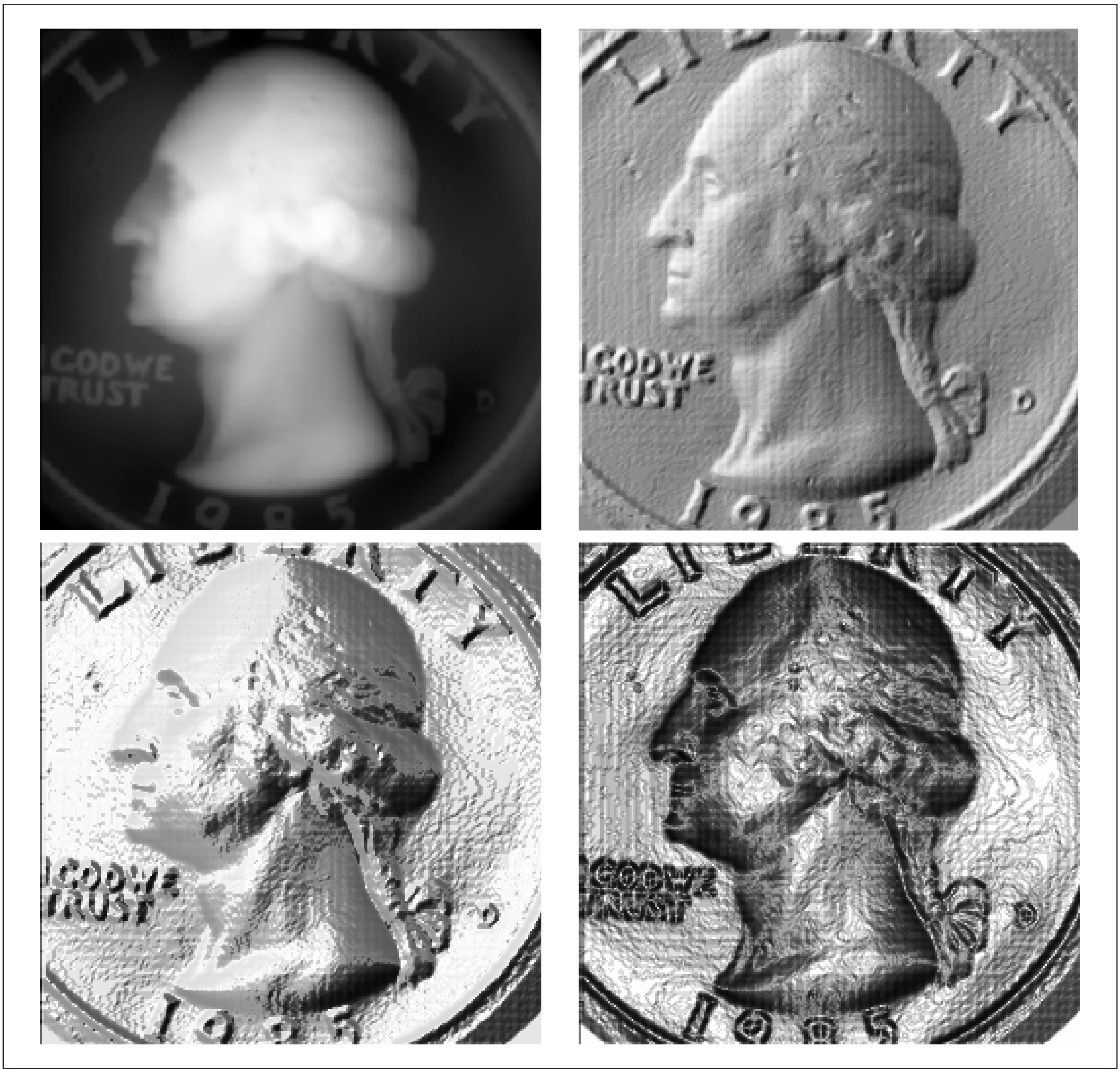


Figure 50. Range image of a coin (produced by a scanned stylus microscope), and three renderings using Phong shading with different specularity and incident light positions.

From a series of images with different light source locations it is possible to interpret the geometry of the object from its appearance. We do this automatically, because our brains have evolved in a world that provides many opportunities to learn about the effect of tilting objects on their appearance, and the effect of coating a surface with different materials.

Changing the appearance of a surface from one material to another, or altering the light source color or position, can help us to notice important details on an object. Partly this is due to enhancement of the reflections from particular features, and partly to violating the expectations we normally have when viewing a surface and consequently forcing our attention to all of the details in the image. There is a powerful demonstration of this at
<http://www.hpl.hp.com/news/2000/oct-dec/3dimaging_files/tablet_specular_VR.html>
where a surface imaging technique developed by Tom Malzbender at Hewlett Packard Labs is shown. A series of images taken with a single, stationary camera but with lighting from many different (known) orientations is used to compute the orientation and albedo of the surface at each location on an object. This data set is then used to render an image of the surface with any characteristics, including those of an ideal metal, while the viewer interactively moves the light source position. The example in Figure 51 shows the recovery of fine details from an ancient clay tablet.

The technique that underlies this calculation is called “shape from shading” or “photometric stereo.” Instead of taking two or more pictures from different viewpoints, as in stereoscopy, photometric stereo uses multiple images from the same viewpoint but with different illuminations. Shape from shading uses the known distributions of diffuse and specular reflec-

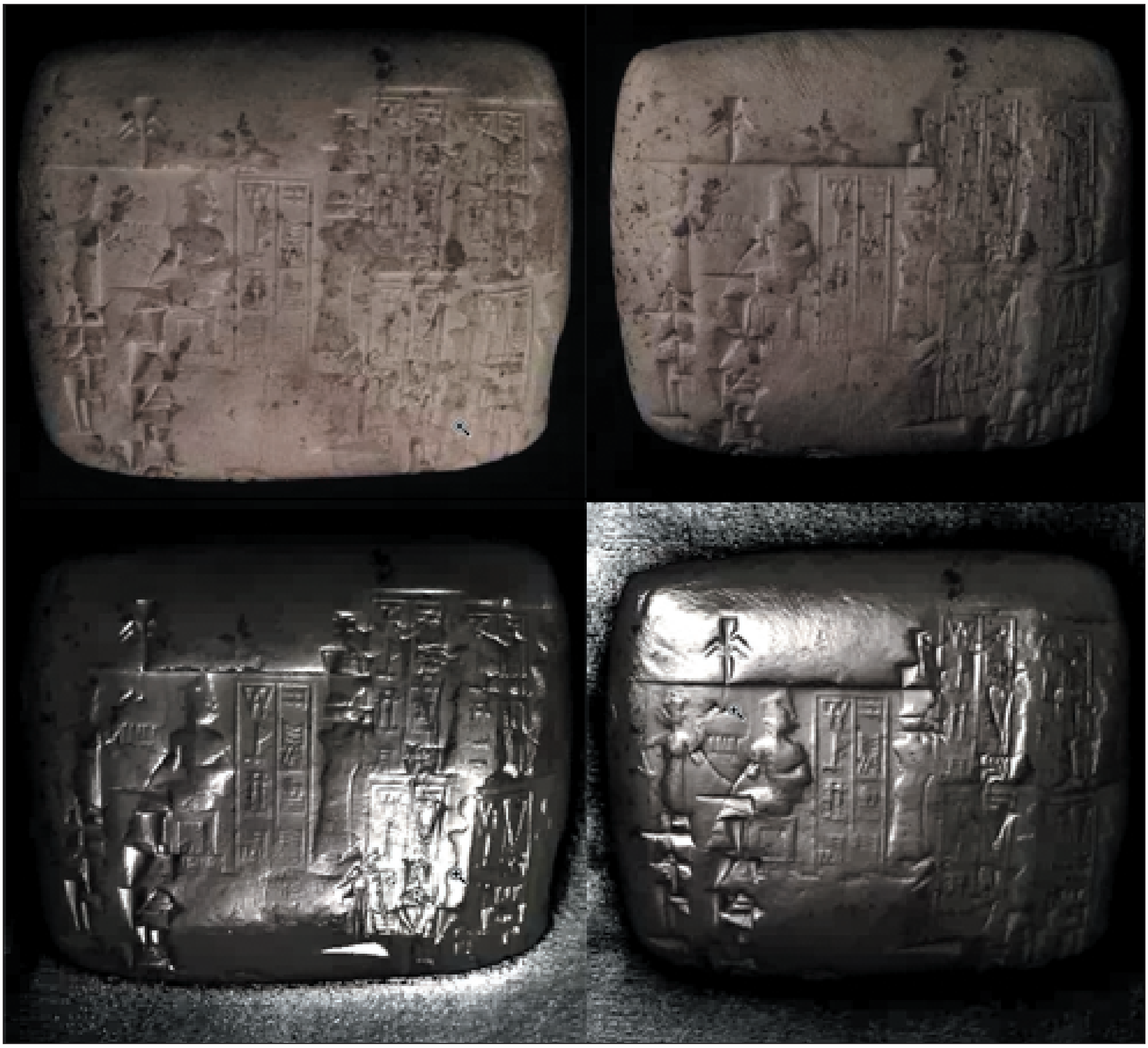


Figure 51. Images of an ancient clay tablet with different illumination positions (top), and rendered as a metallic surface to reveal subtle detail (bottom).

tions for a particular type of surface to estimate changes in the local slope of the surface with respect to the lines of sight from the light source and the viewpoint. The weakness of the shape-from-shading approach is that it deals only with differences in intensity (and hence in slope). Determining the actual surface elevation at each point requires integrating these slope values, with an unknown constant of integration. Nevertheless, the method has numerous applications, and also serves to illustrate a computation that our minds have been trained by years of practical observations to make automatically.

The mental shortcuts that enable us to interpret brightness variations as shape are convenient and often correct (or at least correct enough for purposes such as recognition and range-finding). But they are easily fooled. Surface brightness can change for reasons other than geometry, such as the effect of intentional or unintentional coatings on the surface (e.g., oxidation, stains). There may also be nonuniformities in the illumination, such as shadows on the surface. If these are not recognized and compensated for, they will influence our judgment about the surface geometry.

One thing that we are conditioned to expect from real-world viewing is that lighting comes from above, whether it is the sun in the sky or lights above our desk surface. If that expectation is violated, our built-in shape from shading calculation reaches the wrong conclusion and interprets peaks and pits and vice versa (Figure 52). Such illusions are amusing when we recognize them, but sometimes we may remain fooled.

Images that come from novel modalities, such as the scanning electron microscope, appear to be familiar and readily interpretable because the brightness of surfaces varies with slope, just as in the true shape from shading situation. But different physical processes are involved, the mathematical relationships between brightness and geometry are not quite

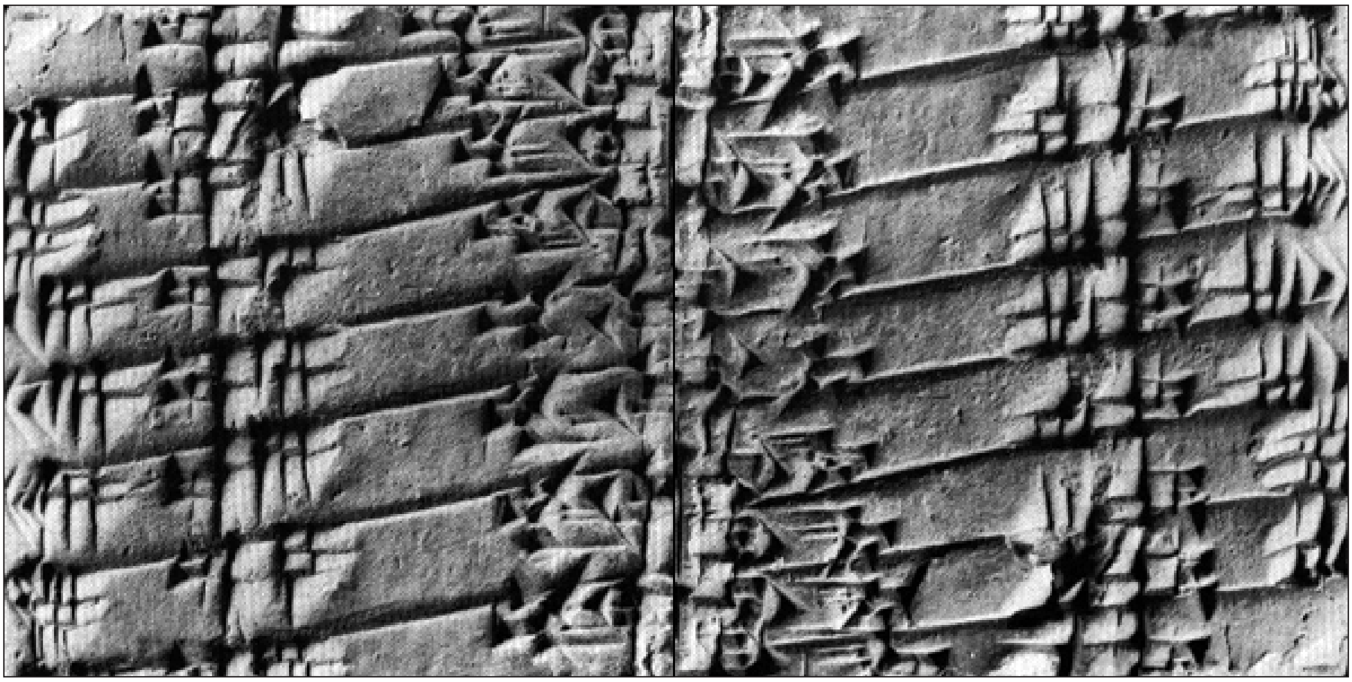


Figure 52. Rotating the same image (of cuneiform indentations in a clay tablet) by 180 degrees makes the pits appear to be peaks.

the same, and misinterpretations can occur. For one thing, the appearance of edges and fine protrusions is very bright in the SEM, which does not occur in normal light scattering from surfaces (Figure 53).

Many other types of images, such as the surface maps produced by the AFM based on various tip-sample interactions, are commonly presented to the viewer as rendered surface representations. These are typically generated using the strength of a measured signal as a measure of actual surface geometry, which it may not be. Electronic or chemical effects become "visible" as though they were physical elevations or depressions of the surface. This is an aid to "visualization" of the effects, taking advantage of our ability to interpret surface images, but it is important (and sometimes difficult) to remember that it isn't really geometry that is represented but some other, more abstract property.

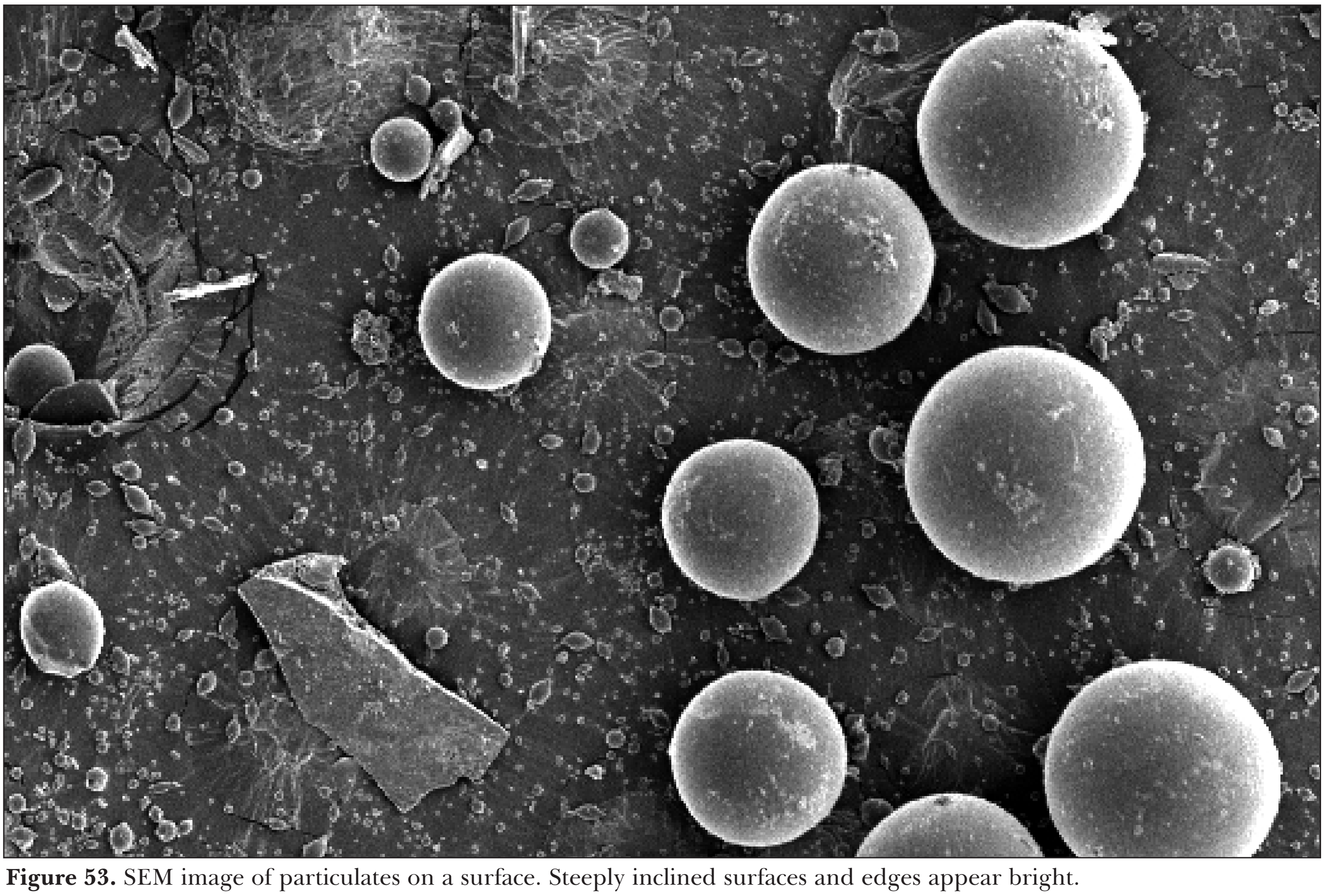
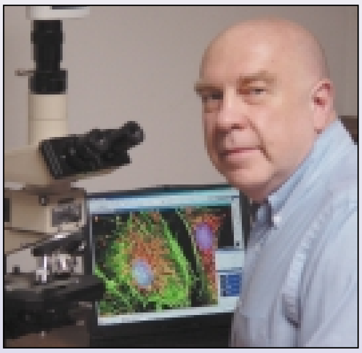


Figure 53. SEM image of particulates on a surface. Steeply inclined surfaces and edges appear bright.

Article



John Russ is the author of The Image Processing Handbook, Computer Assisted Microscopy, Practical Stereology, Forensic Uses of Digital Imaging, Image Analysis of Food Structure, as well as many other books and papers. He has been involved in the use of a wide variety of microscopy techniques and the computerized analysis of microstructural images for nearly 50 years. One of the original founders of Edax International (manufacturer of X-ray analytical systems), and the past Research Director of Rank Taylor Hobson (manufacturer of precision metrology instruments), he has been since 1979 a professor in the Materials Science department at North Carolina State University. Now retired, he continues to write and lecture on topics related to image analysis.

"The first section of this three-part paper has emphasized the dependence of vision on local comparisons of brightness, color, orientation, and feature relationships. In the next part, comparisons over time are included to interpret motion, and comparisons over longer ranges are shown to influence judgments of distance. In addition, the tendency of people to see only a few things in a scene, and to see what they expect to see in a given context, is illustrated.

The third and concluding part deals with object recognition"

Seeing the Scientific Image, Part 3

John C. Russ

Materials Science and Engineering Department,
North Carolina State University, Raleigh, NC

Size matters

The size of features is determined by the location of the feature boundaries. The only problem with that rather obvious statement is deciding where the boundary lies. Human vision works by finding many types of lines in images, which include edges of features, based on locating places where brightness or color changes abruptly. Those lines are treated as a sketch (called the "primal sketch"). Cartoons work because the drawn lines substitute directly for the edges that would be extracted from a real scene.

Using a computer program to extract edge lines (Figure 54) illustrates the idea of the sketch. The computer program finds the location (to the nearest pixel) where the maximum change in brightness occurs. The sketch extracted by the retina isn't quite the same. For one thing, gradual changes in brightness are not as visible as abrupt changes, and the change must be at least several percent to be noticed at all. Changes in color are not as precisely located, and some color changes are much more noticeable than others (variations in the green part of the spectrum are noticed least).

Furthermore, people do not interpret the edge line in a consistent way. A simple demonstration can be found in the way we cut out patterns, and there is some indication that there is a sex-linked behavior involved. Girls cutting cloth to make a garment tend to cut outside the line (better seams too wide than too narrow); boys cutting out model airplane parts tend to cut inside the line (so parts will fit together). The same habits carry over to tracing features for computer measurement. A trivial difference, perhaps, but it raises the interesting question "Is the edge a part of the feature or a part of the surroundings?" In many cases, that depends on whether the feature is dark on a light background (in which case the edge is likely to be seen as part of the feature) or the converse.

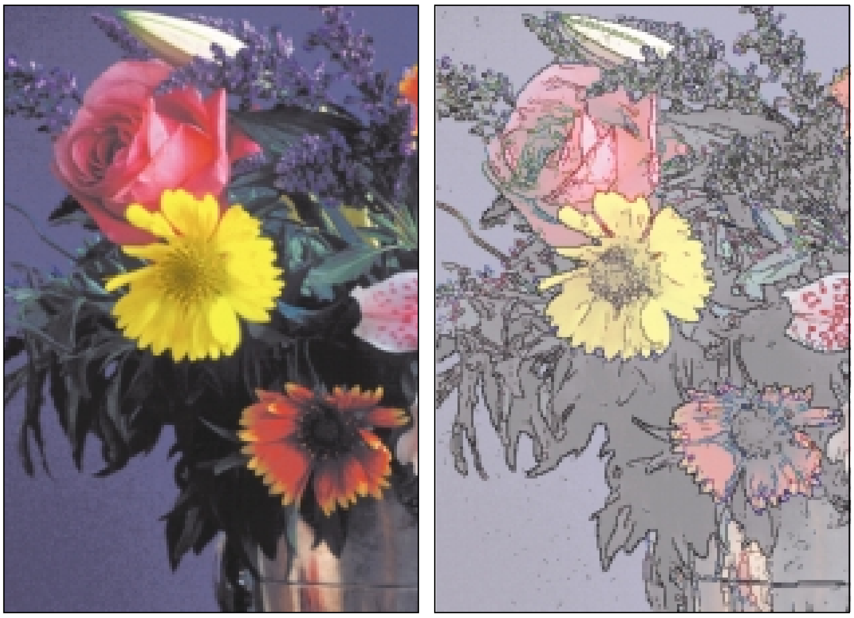


Figure 54. Extraction of the edges from an image produces the primal sketch of the scene.

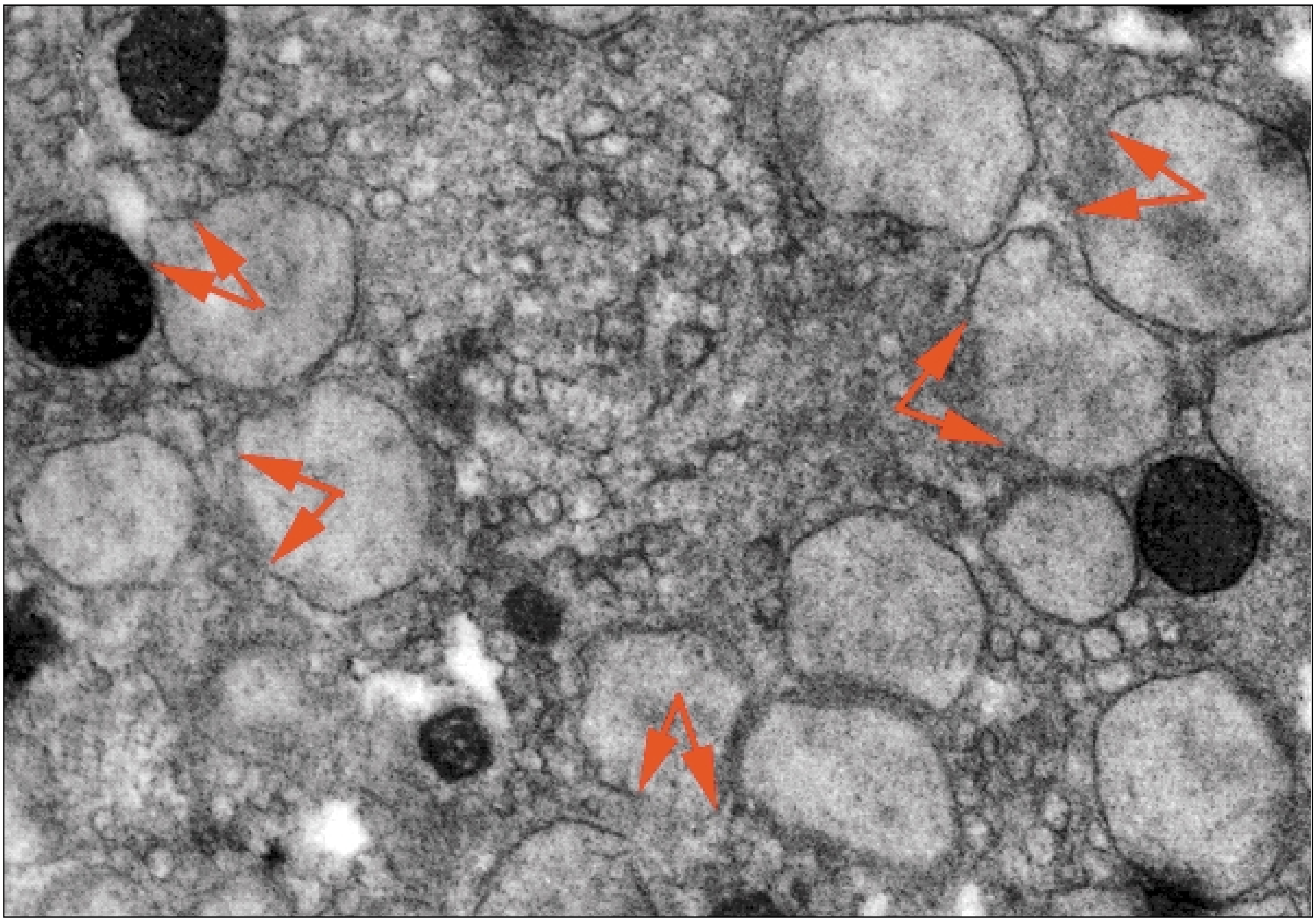


Figure 55. TEM image of stained tissue. The membrane boundaries of the organelles are not visible in some locations (arrows) but human vision "knows" they continue and completes them with simple smooth curves.

In many real images, the boundaries of features are not uniform. Variations in brightness and contrast cause variation in judgment of the location of the feature edge, and hence in its size and shape. In some cases, the boundary disappears in places (Figure 55). Human vision is not bothered by such gaps (although computer measurement certainly is). We fill in the gaps with simple, smooth curves that may or may not correspond to the actual shape of the feature.

Boundaries are certainly important, but there is evidence that features are represented conceptually not as a collection of boundaries but as a simplified midline. The pipe-cleaner animals shown in Figure 56 are recognizable because we fill out the bodies from the "skeleton" shown. This is not the actual skeleton of bones, of course, but the one used in computer-based image analysis, a set of midlines sometimes called the medial axis of the object. The topology of the skeleton (the number of branches, ends and loops) provides critical information for feature recognition by humans and machines.

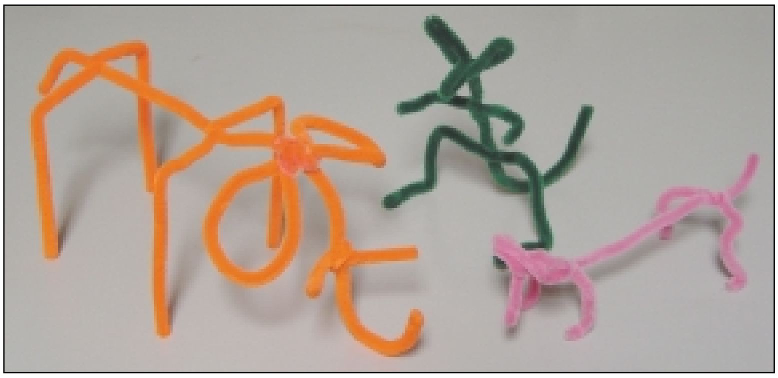


Figure 56. Pipe-cleaner animals (elephant, kangaroo and dachshund) represent solid objects by their skeletons.

Whether the boundaries or the skeletons of features are used for representation, comparisons of the size of features are strongly affected by their shape, position and brightness. A map of the continental United States illustrates this well (Figure 57). In order to compare the sizes of two states we literally drag the image of one onto the other, in our minds. Since the shapes are different, the fit is imperfect. How the parts that "stick out" are treated depends on their perceived importance. For example, comparing Oklahoma to Missouri is tricky because the panhandle of Oklahoma is pretty skinny and easily overlooked (but Oklahoma is larger than Missouri).

Florida is about the same size as Wisconsin, but they are different colors and far apart and comparison is very difficult. Colorado has a simple rectangular shape, difficult to compare to Nevada or Oregon which are not so regular and tend to appear smaller. The greater vertical extent of North Dakota is visually impor-

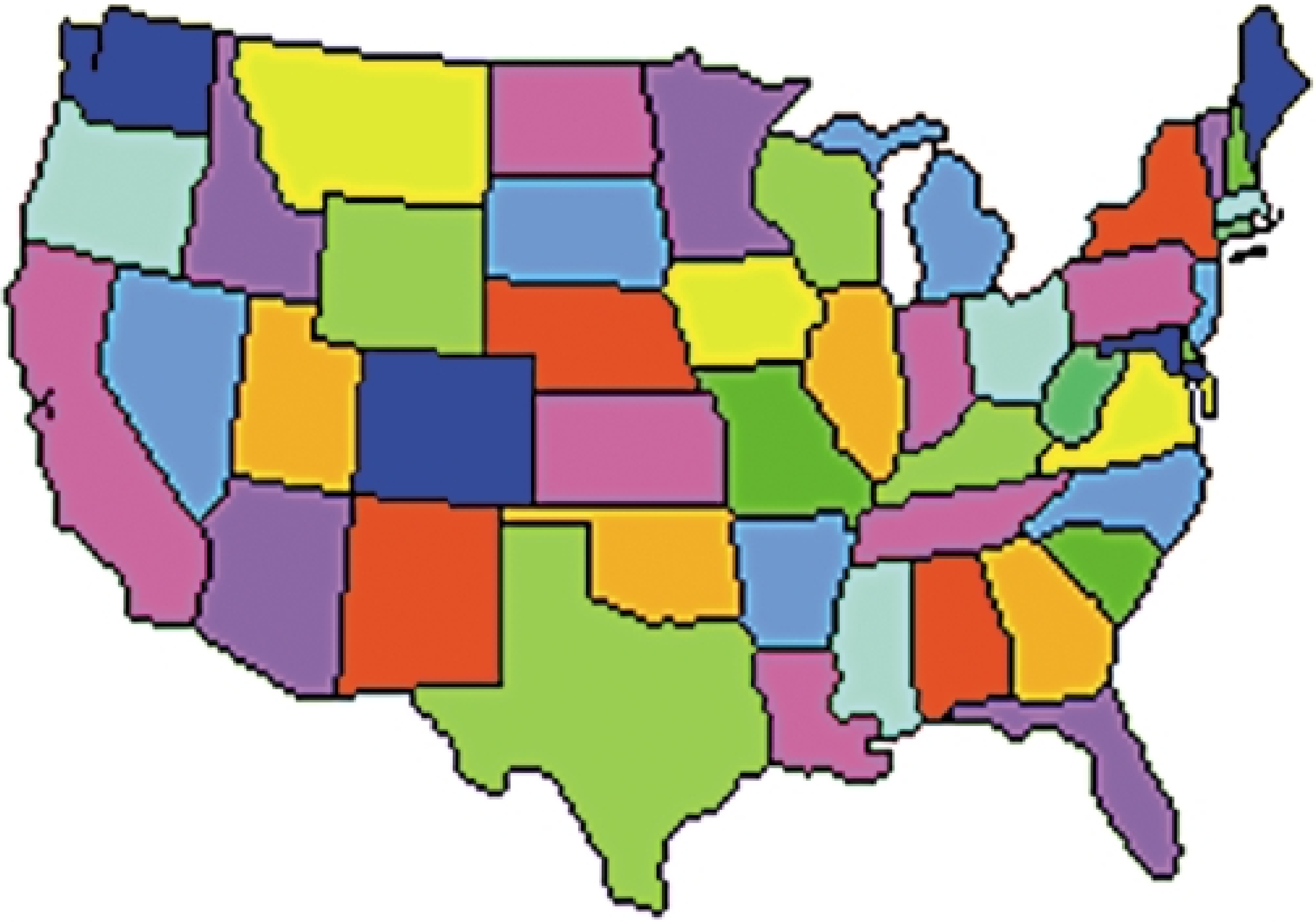


Figure 57. The continental United States.

tant and leads to the erroneous conclusion that the state is larger than South Dakota. Vertical extent is important in comparing Illinois to Iowa and New York, as well, as are the differences in shape and color (and the fact that New York is far away).

Visual judgments of size are very error prone under the best of circumstances, and easily swayed by seemingly minor factors, several of which have been illustrated in these examples. Another very common mistaken judgment of size involves the moon. Most people report that it appears to be larger by one-third to one-half when near the horizon that when high in the sky (Figure 58), probably because at the horizon there are other structures which the eye can use for comparison. Vertical extent is generally considered more

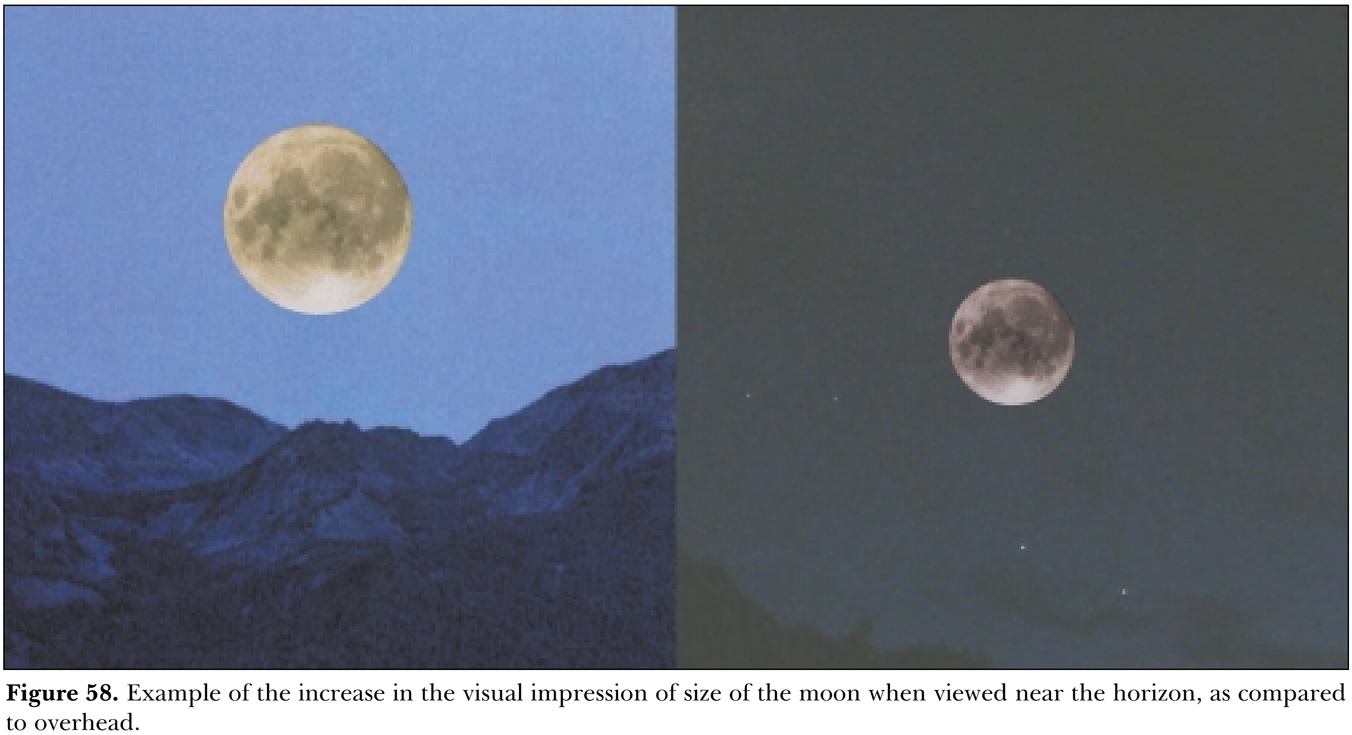


Figure 58. Example of the increase in the visual impression of size of the moon when viewed near the horizon, as compared to overhead.

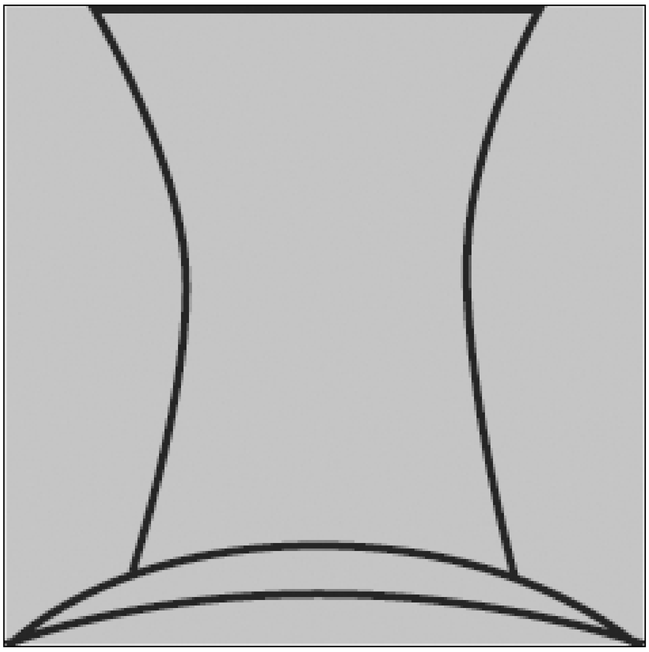


Figure 59. The "top hat" illusion: In this exaggerated drawing of a top hat, the height appears to be much greater than the width, but in fact they are exactly the same.

important than horizontal extent (Figure 59). Features that contrast more with their surroundings are generally considered to be larger than ones with less contrast. Departures from geometrically simple shapes tend to be ignored in judging size.

Also, the context of the scene is also very important. In the discussion of stereoscopic vision and interpretation of the third dimension it was noted that expectation of constant size is one of the cues used to judge distance. It works the other way, as well. We expect the rules of perspective to apply, so it one feature is higher in the scene than another, and is expected to be resting on the ground, then it is probably farther away, and thus it should appear to be smaller. If it is actually the same size in the image, we would tend to judge it as being larger in actuality. Unfortunately, when viewing images for which the rules of perspective do not apply this "correction" can lead to the wrong conclusions.

Shape (whatever that means)

Shape is extremely important in visual recognition of objects. Even if the size or color of something is changed radically (a miniature pink elephant, perhaps), the shape provides the important information that triggers identification. But what is shape? There are very few common adjectives in English or any other language that really describe shape. We have plenty for size and color, but few for shape. Instead, we describe shape by saying that something is shaped "like an elephant" - in other words we don't describe the shape at all but simply refer to a representative object and hope that the listener has the same mental image or model that we do, and identifies the same important shape features that we do. The few apparent exceptions to this - adjectives like "round" - actually fall into this same category. Round means "like a circle" and probably everyone knows what a circle looks like.

But what does it mean to depart from being round like a circle? Unfortunately there are lots of ways to become less like a circle. Figure 60 shows just two of them: one feature has been stretched horizontally but is still smooth while the other has remained equiaxed but with a rippled edge. Many other variations with jagged edges, stretching in more or other directions, and so forth are possible. Which of these features should be considered "rounder?"

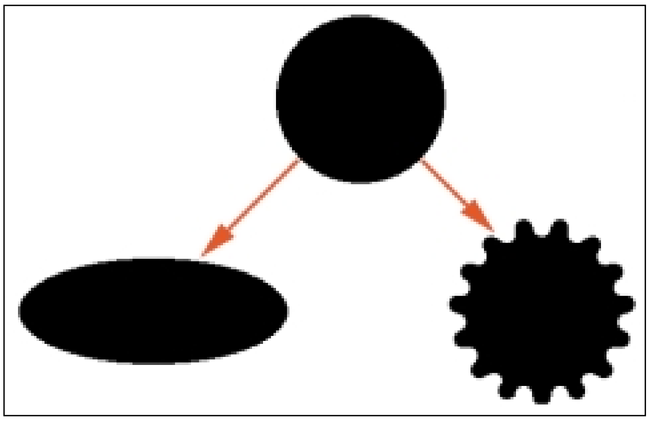


Figure 60. The two lower shapes have the same area as the circle but are not circular.

When computers measure features in images, there are a lot of mathematical ways that they can describe shape. The most commonly used are simple dimensionless ratios of size measurements. There are a lot of ways to measure size (e.g., area - with or without holes, perimeter - total or exterior only, maximum and minimum caliper dimension, equivalent circular diameter, radius of the smallest circumscribed or largest inscribed circle, area and perimeter of the convex hull or taut string boundary, to list only the most common), and most of these are easily understood by humans because they use familiar words for familiar concepts. But these size parameters can be combined in an almost limitless number of ways to produce formally dimensionless shape parameters. Two of the more widely used are

$$\mathbf{4\,\pi\,Area\,/\,Perimeter^2}$$

$$\mathbf{4\,Area\,/\,\pi\,Length^2}$$

The first of these two shape parameters uses area and perimeter and is generally sensitive to the departure from roundness exhibited by the feature on the right, while the second one uses area and maximum caliper dimension, or length, and is generally insensitive to changes in the smoothness of the perimeter but does vary with elongation.

The names assigned to these are completely arbitrary and have no familiar intrinsic meaning. Furthermore,

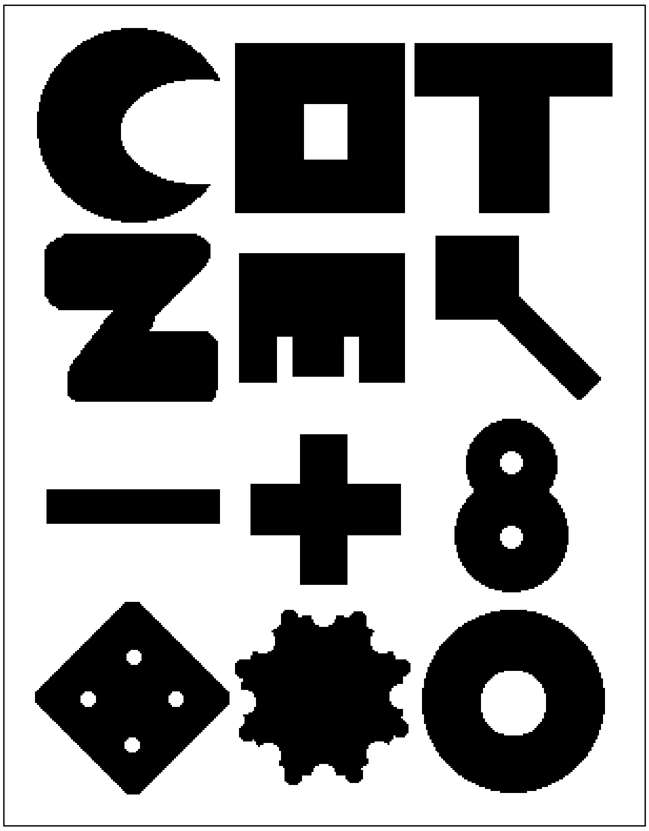


Figure 61. Twelve visually different shape with identical values of the shape parameter

4 π Area / Perimeter^2

the various writers of software for image measurement use the names differently. Some call the first one (or its inverse) roundness; some call it formfactor and use roundness for the second one, or use the word circularity. Some use other formulas and other names. Without an accepted human meaning for these concepts no consistency should be expected, and indeed none is found.

An important additional problem with these dimensionless ratios is that they are not at all unique. It is possible to construct an infinite number of objects whose shapes appear very different to a human but which have identical values of formfactor, as shown in Figure 61.

Still, these shape factors can be very useful for computer recognition purposes. Combined with one other type of shape descriptor (the number of holes, a fundamental topological property that will be discussed below), they often allow expert systems to be established to perform machine vision chores. As a trivial example, consider the task of recognizing the letters A through E as shown in Figure 62. Although the letters use different fonts (serif and sans-serif), are printed in different sizes and arbitrary orientations, humans have no difficulty in identifying them. Machines can do this too, but in a different way. Figure 62 shows a flow chart for an expert system that uses several shape descriptors to accomplish it.

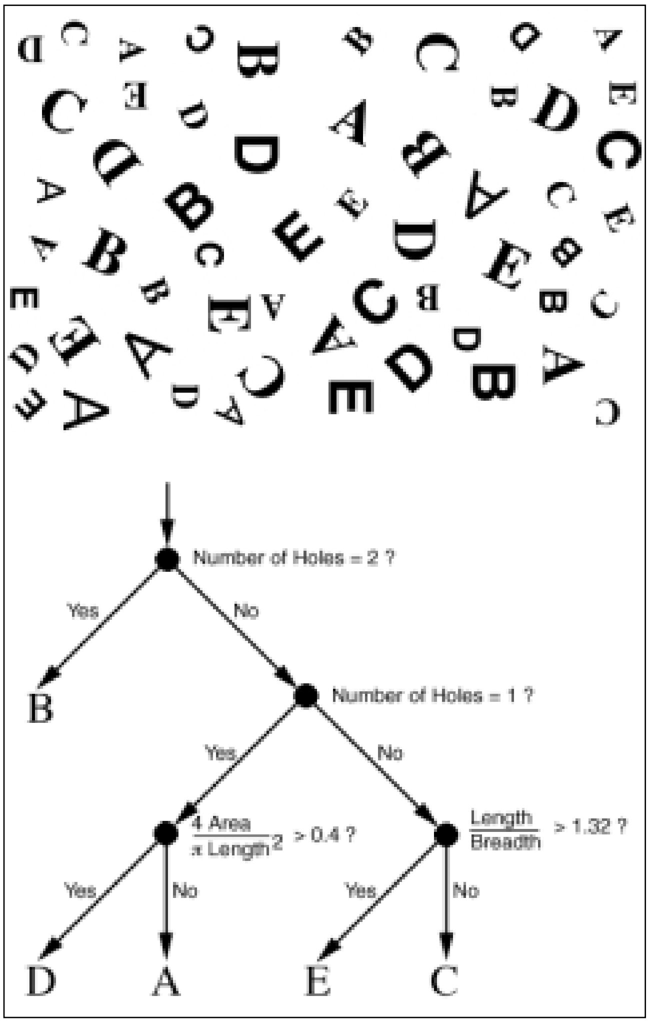


Figure 62. Some recognizable shapes and a flow chart that can identify them.

This kind of procedure is very brittle. Adding more characters or extreme fonts that are still easily recognized by a human breaks the procedure and an entirely new one may need to be devised. It is interesting that our familiarity with the Roman alphabet allows us to identify the characters very readily even if they are presented in an unfamiliar way (e.g., rotated). When presented with less familiar characters (unless you happen to read Hebrew), a human takes much more time to make the comparisons and determine how many different identical characters are present, and which is which, as shown in Figure 63. This indicates that for unfamiliar shapes the comparison is performed in the usual way by mentally dragging and rotating the characters so they can be compared directly, while in the case of the familiar shapes the labels are applied immediately and it is only the labels that are matched, and not the actual shapes that must be compared.

One consequence of this behavior for scientific observations is that the researcher who has (by dint of long hours of looking, or based on other knowledge) become sufficiently familiar with a class of objects to recognize them, even if he or she does not formally have a list of identifying characteristics, and even if

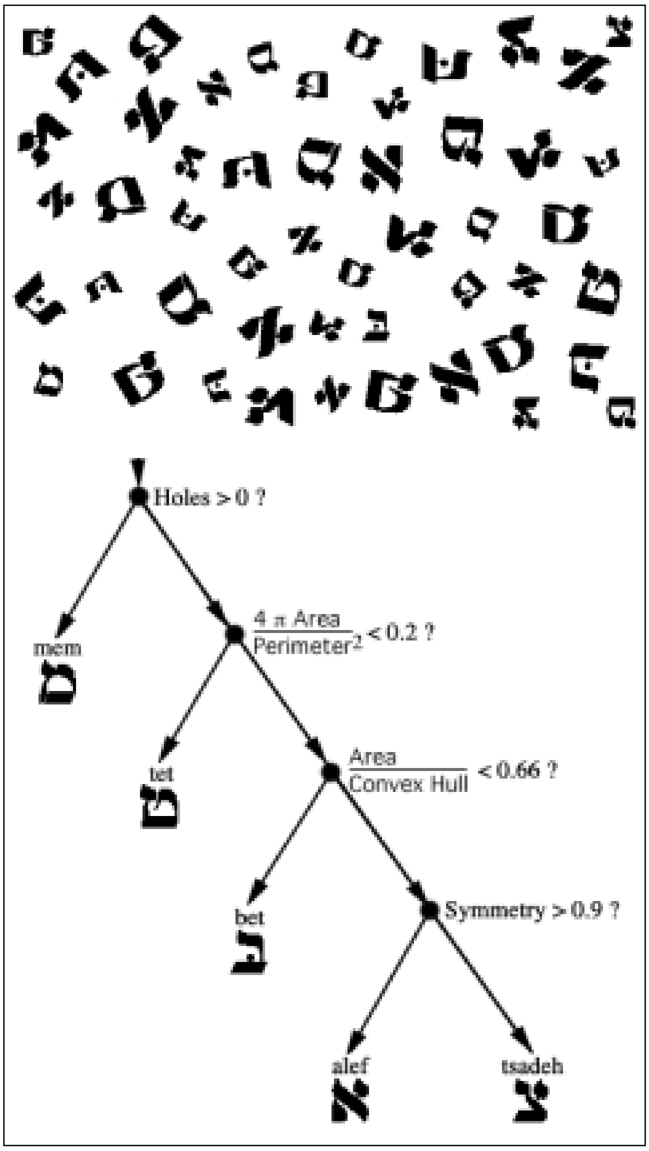


Figure 63. Flow chart for identification of some Hebrew letters. Symmetry is defined here as one minus the ratio of the distance from the feature centroid to the geometric center (center of a circumscribed circle) divided by the radius of that circle; it is 1.0 for a perfectly symmetrical feature.

that recognition is flawed, will immediately apply labels to the features in an image, whereas someone without that level of familiarity may struggle to make the same identification or comparison. If the labeling criteria used are imperfect, false positive or negative identifications will occur for the experienced viewer but not for the novice, who is forced to make a more detailed comparison.

These dimensionless ratios are useful shape factors for computer identification because there are so many of them available and they require so little computational effort, but they do not correspond to what humans call shape, and their use does not mimic the way that people perform feature recognition.

There are other computer-based ways to describe shape. One of the most computationally intensive is to "unroll" the feature periphery as a plot of radius versus angle, or of slope versus distance along the boundary, and then perform a Fourier transform on it. The first few dozen coefficients in the Fourier expansion of the plot are then used in statistical classification procedures that have demonstrated great power in a few areas of application such as sedimentology and palynology. The method often accomplishes robust identification in cases where humans are confused by the apparent wide variations in perceived shape, by finding some underlying fundamental characteristics that can be statistically extracted from the clutter. But people clearly do not "see" the same characteristics that allow this technique to identify the sediment deposited by two glacial rivers based on the 5th and 8th Fourier coefficients. It is the lack of correspondence between what these admittedly powerful statistical methods can accomplish and the ability of a human observer to extract similar information from the images that has limited its application.

Topology and Fractal Dimension

It seems that instead of these numerical properties of shape, people rely primarily on two specific kinds of information for shape recognition. Fractal dimension will be discussed below. The other principal kind of information is topological and is best illustrated by using a computer processing operation to reduce a shape to its skeleton. The skeleton, mentioned previously, is the midline of the feature, produced by an iterative removal of pixels from the boundary until the only ones left cannot be removed without breaking the feature into pieces. As shown in Figure 64, the end points and branch points of this skeleton correspond to the basic topological features of the original shape, and they can be very easily located because end points are pixels that have only one neighbor and branch points (or nodes) have more than two. Euler figured out long ago that these topological features obey the relationship

Number of Holes (or Loops) = Number of Branches - Number of Ends - Number of Nodes + 1

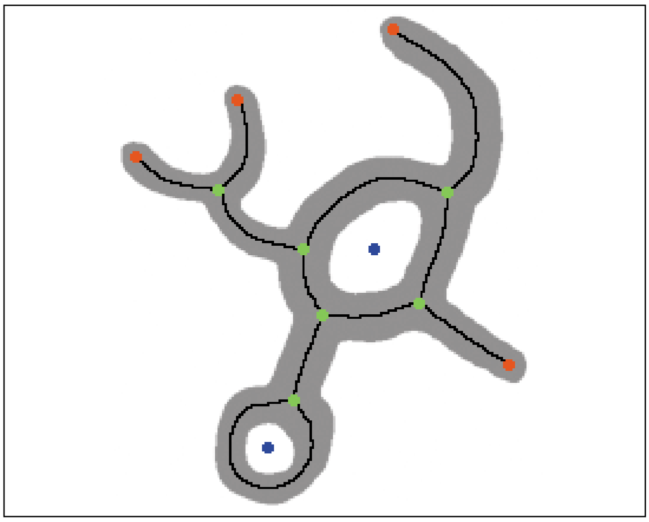


Figure 64. An arbitrary shape (grey) with its superimposed skeleton (black). The 4 end points (red), 6 branch points (green) and 2 loops (blue) are marked.

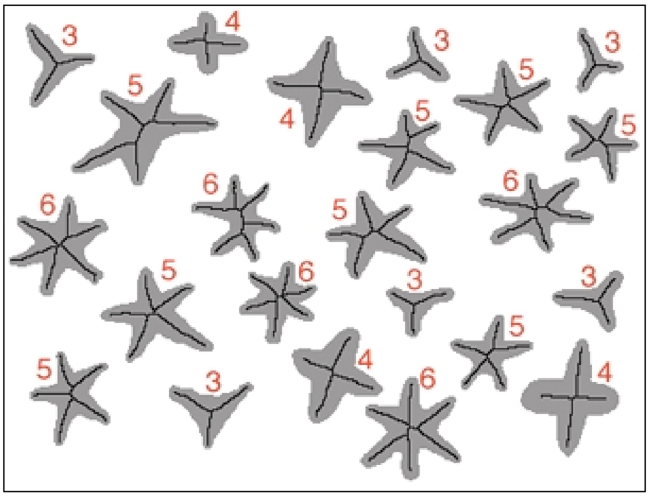


Figure 65. Star shapes with superimposed skeletons and labels with the number of end points.

The efficiency of using the skeleton for measurement of these topological features is illustrated in Figure 65. Recognition of the basic shape of the different stars (according to whether they have 3, 4, 5 or 6 points) is immediate for a human. By counting the number of end points the computer can label each one with that same topological property.

Instant recognition of the number of end points becomes more difficult for humans when the number is large or when the shapes are more complex and have other variable (such as the length or thickness of the arms) as well. Figure 66 shows an example of the latter case in which the arms must be counted, one by one, for some of the features. People are not good at counting, and often make errors.

When the number of end points becomes large, gestalt counting no longer works (for various individuals this usually happens somewhere in the range of 6 to 12). People don't count things very well. Visually counting the number of teeth on the gear in Figure 67 is slow and error-prone, but the computer method using the skeleton end points is still fast and robust.

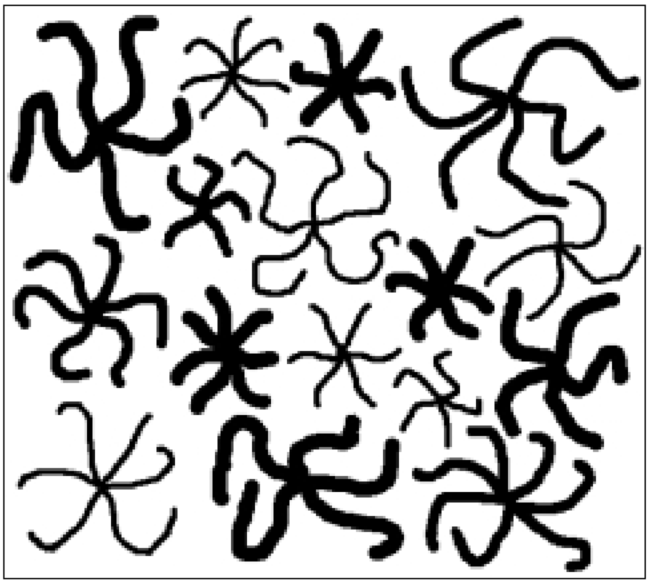


Figure 66. More complex star shapes with five, six and seven arms.

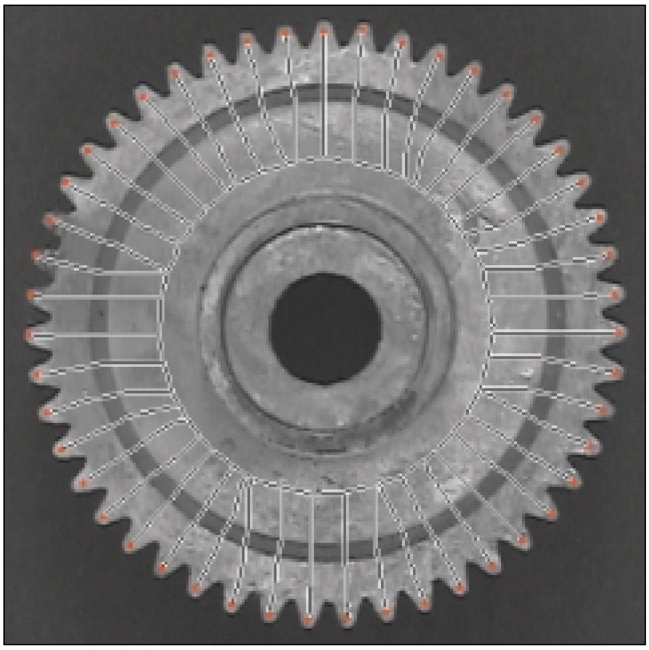


Figure 67. Image of a gear with superimposed skeleton whose end points mark the 47 teeth.

Human vision apparently uses topological information and something like the skeleton for shape characterization. The key topological features - end points, branch points, corners and loops - are extract-

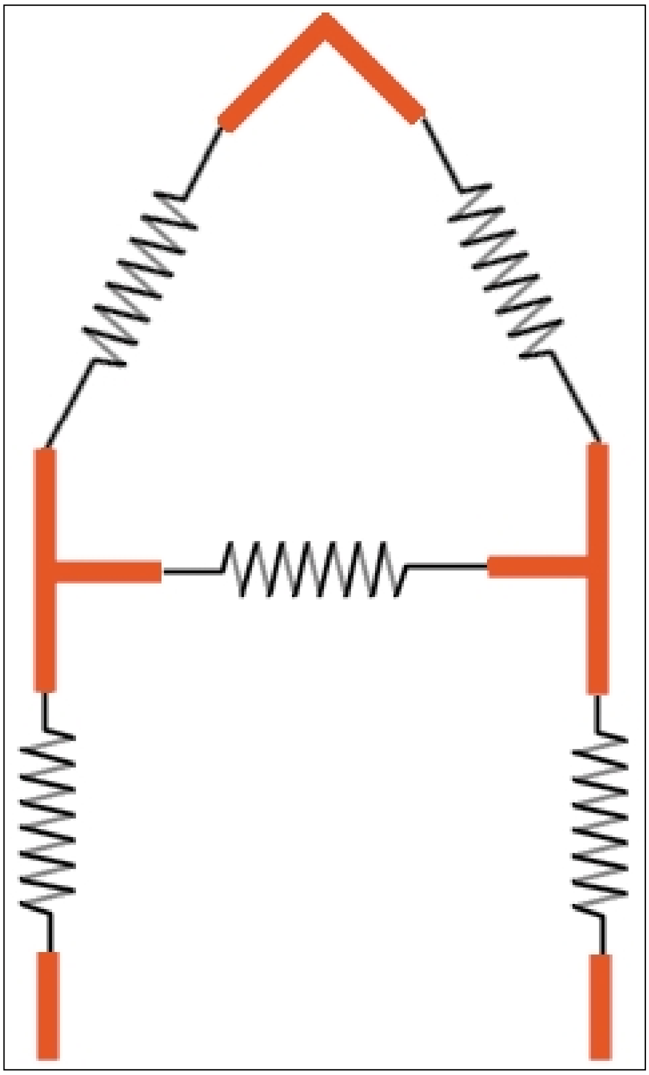


Figure 68. Key topological features of the letter "A."

ed along with their pattern of connection. But, as shown in Figure 68, the distances between of those connections are not so important. For the letter "A" as an example, the two branch points, two downward pointing end points, and the sharp corner at the top can considered to be connected by springs. Any distortion of the figure that stretches the springs but keeps the basic order of connections correct will be recognized as an "A" and in fact is sufficient to distinguish it from any other letter in the alphabet.

To perform the same task described above of identifying the letters A through E, the key topological features of each are identified as shown in Figure 69. Note that this method is very robust to changes in the size of the letters and can accommodate much greater changes in fonts or even handwriting than the method using dimensionless ratios shown above.

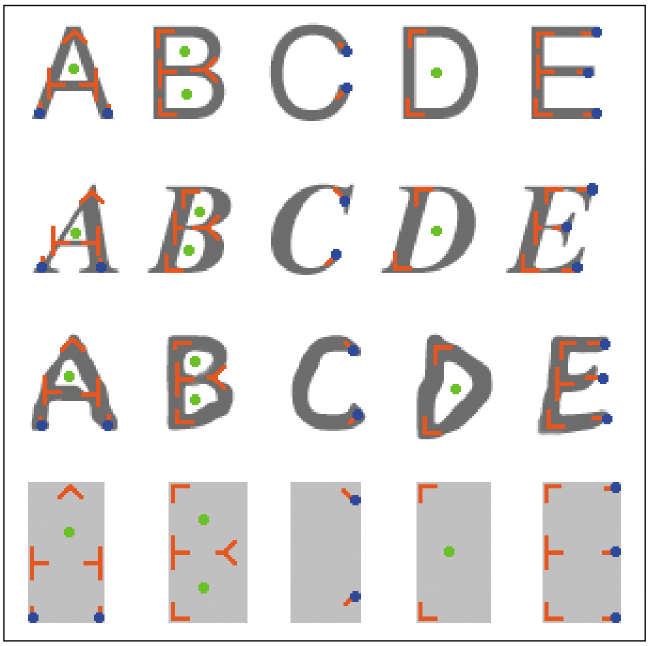


Figure 69. Key topological features of the letters A through E.

It is the same ability to use just a few key features (corners, ends and branch points) to characterize feature shape that accounts for one of the common illusions. Kanisza's triangle (Figure 70) is constructed in the mind by linking together the three well defined corner points. The linking is always done with smooth, although not necessarily straight lines. As for the case of the end points in a star, the human ability to recognize the gestalt of polygons works best with a small number of sides. Once the shape has been formed in the mind, most people report that the interior of the triangle (which is of course the illusion) appears to be brighter than the background upon which it "rests."

Once the basic topological form of the object shape has been established, the second property of shape that people seem to instinctively recognize is a measure of the smoothness or irregularity of the boundary. Because so much of the natural world has a geometry that is fractal rather than Euclidean, this takes the form of an ability to detect differences in boundary fractal dimension.

So what is a boundary fractal dimension? For about the last hundred years mathematicians have been intrigued by curves that have the unusual, even disturbing property of an undefined length. Examining the line at ever finer detail simply reveals more irregularities and more length. In terms of examining scientific objects, whether it is the margin of a leaf or the fracture of a metal, increasing the image magnification reveals more detail in a hierarchy that extends over many decades of scale. Not everything has this "self-similar" property. Fluids with surface tension, cells with elastic membranes, and most man-made surfaces have a well defined Euclidean boundary whose length can be defined and does not change with magnification. Such boundaries for objects on two-dimensional images have the topological dimension of a line, exactly one.

But irregular boundaries that are fractal have a length that increase regularly with magnification. The classic illustration is Richardson's "How Long is the Coast of England?" example. Measuring that length on a map using dividers set at 10 km will produce one result. Walking along the coast with a 100 meter chain would follow more of the irregularities and produce a larger measurement. Crawling along the shore with a meter stick would result in a further increase, and so on. Plotting the length values versus the length of the measuring tool, on log-log axes, gives a straight line whose slope M equals 2 - D where D is the fractal dimension of the boundary, a number between 1.0 and 1.999. The higher the number the more irregular, or rough the boundary is perceived to be. In the Hawaiian islands, for example, the older (and more weathered) islands have a dimension greater than the younger (and smoother) ones.

Many natural structures have been shown to have this fractal nature, with typical values in the range up to about 1.35, and apparently people, while they certainly do not measure the dimension, have a very robust ability to compare features and put them into the correct order in terms of boundary "roughness." As usual, human vision compares things best when

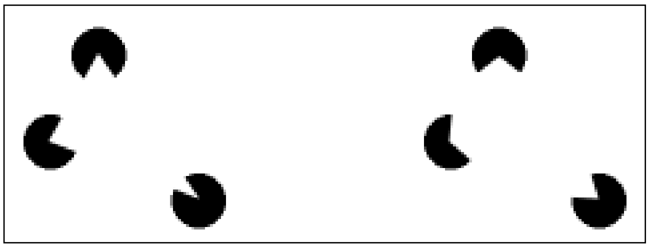


Figure 70. Kanisza's triangle is an illusory region visually formed by linking corner markers with straight lines or gentle curves.

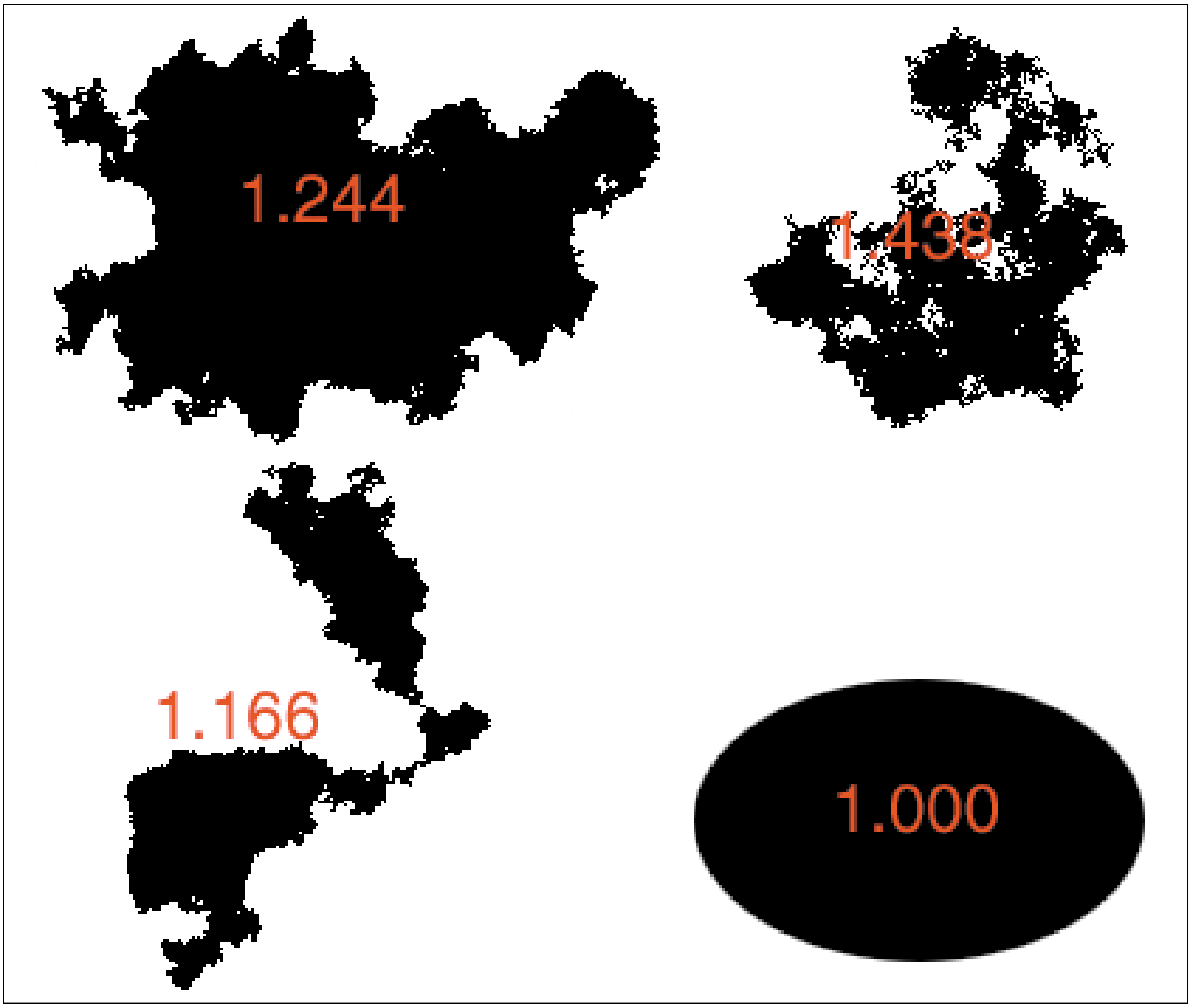


Figure 71. Several shapes with their fractal dimensions as measured by the computer.

they can be placed side by side in the same field of view, performs the comparison between two features at a time, and only painstakingly constructs a ranking for a field of objects. It also suffers when comparing a currently viewed object to one from memory, because the fine details needed to represent boundary irregularity are typically either not remembered very well or in some cases recalled with too much emphasis. Computer measurement of fractal dimension can also be performed; the most accurate method constructs a plot of the number of pixels as a function of their distance from the boundary, whose slope gives the dimension. As shown by the examples in Figure 71, the feature's fractal dimension relates only to this boundary irregularity and not to other aspects of shape such as the topology or to ratios such as length / breadth.

Since the visual recognition of shape is primarily based on simple topological form and boundary irregularity, it might be useful to employ those parameters to communicate a description of shape from one person who is familiar with a class of objects to another, who is not. The mathematical descriptions of the parameters are perhaps unfamiliar but not really difficult, and the ideas at least are comfortable. Furthermore, computers can be easily programmed to extract the same information. In fact they are: it is the topological shape of printed characters that is used in most optical character recognition (OCR) software that converts pages of printed text back into an editable computer file.

So far, however, that approach has been taken only rarely. The most widely accepted method for communicating the information about object shape characteristics that are used for recognition remains showing a picture of the object to another person. For features that are exactly alike, or so nearly alike that the only variations are essentially invisible at the scale of normal viewing, that works fine. Unfortunately in most of the sciences the objects of interest, whether they are defects in a material, cancerous cells on a slide, or a new species of bug, are not identical. The natural variation is significant, although the clues to recognition (if they are correctly chosen) remain present (although not necessarily visible in every image).

In presenting the "representative image" the scientist attempts to communicate these clues to colleagues. the picture almost always needs an extensive supplement in words (think of Arlo's song again - the "Twenty-seven 8x10 color glossy pictures with circles and arrows and a paragraph on the back of each one"). But if they are not as familiar with the objects (which is of course the reason for the communication), will they be able to pick out the same features are clues? And does the selected image adequately represent the range of variation in the natural objects? These dangers are always present in the use of "typical" images even if the image really does give a fair representation, and in most cases it should probably be admitted that the picture was selected not after analysis showed it to be representative in any statistical sense but because the picture satisfied some other, unspoken aesthetic criterion (or worse, was the only good quality image that could be obtained). Anecdotal evidence, which is all that a single picture can ever provide, is risky at best and misleading at worst, and should be avoided if it is possible to obtain and present quantitative data.

Context

Recognition of features is often influenced by the context in which they appear. Sometimes this context is supplied by the image itself, but more often it arises from prior knowledge or independent information. In Figure 72, there are several very different representations of the number five, including ones in languages that we may not know. But once the concept of "fiveness" is accepted, the various representations all become understood. Quite a bit of knowledge and experience that has nothing to do with images is involved in this process, and it happens at higher levels of conscious thought than basic shape recognition. Knowing that pentagons have five sides may help us translate the Greek, or recalling a Cinco de Mayo party may help with the Spanish, and so on.

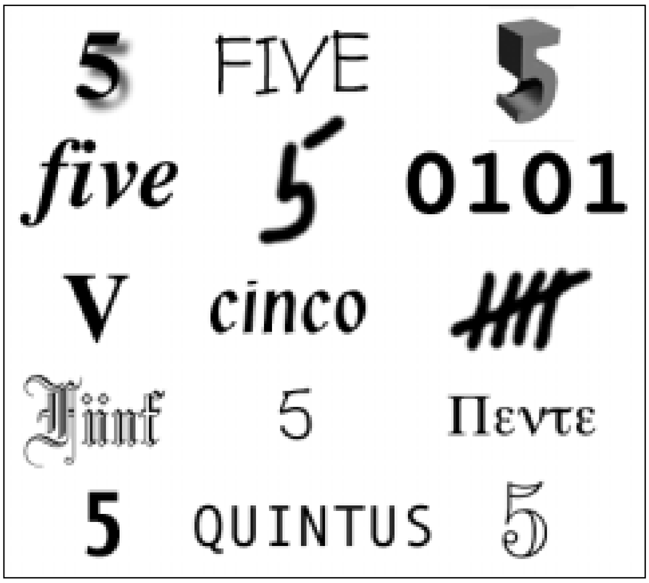


Figure 72. Various representations of "five."

An interesting insight into this recognition process comes from the study of patients who suffer from synesthesia, a phenomenon in which some kind of cross-wiring in the brain confuses the output from one sense with another. People with synesthesia may report that particular notes played on the piano trigger the sensation of specific tastes, for example. In one of the most common forms of synesthesia, looking at a number produces the sensation of a specific color. For instance, in a printed array of black numbers, the fives may all appear red while the threes are blue (Figure 73), and the ability to detect and count the features is extremely rapid compared to the need to identify and count the features consciously.

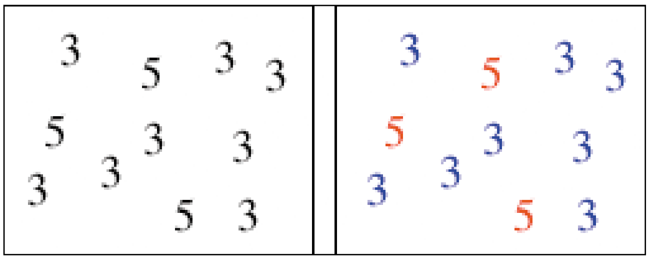


Figure 73. Synesthesia may associate specific colors with numbers, so that they "pop out" of an image.

This cross-activation of different sensory pathways in the brain occurs well before the information rises to conscious levels. The alternative representations of "five-ness" shown above do not trigger these colors. Even modest distortions of the printed number, such as unusual or outline fonts, may be enough to prevent it. The study of these types of brain mis-functions is important for an understanding of the processing of sensory information, extraction of abstract concepts, and formation of connections between seemingly separate areas of knowledge that can occur at subconscious and conscious levels. In this instance, it shows that basic shape recognition happens long before the labels, with their symantic content, are applied to features.

It is even possible to have multiple contexts within the same image. This happens particularly with reading words. We tolerate misspellings and sloppy handwriting because there is usually enough redundancy and context to allow the message to be comprehended even when the image itself is wrong or ambiguous, as shown in the example of Figure 74.

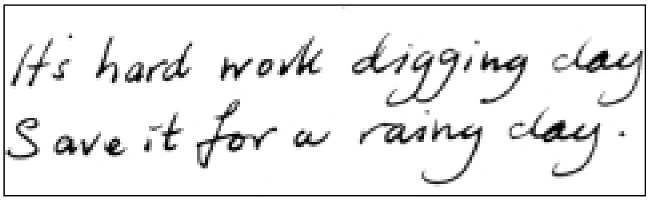


Figure 74. The final words in each line are identical in shape, but can be read correctly because of the context established by the other words.

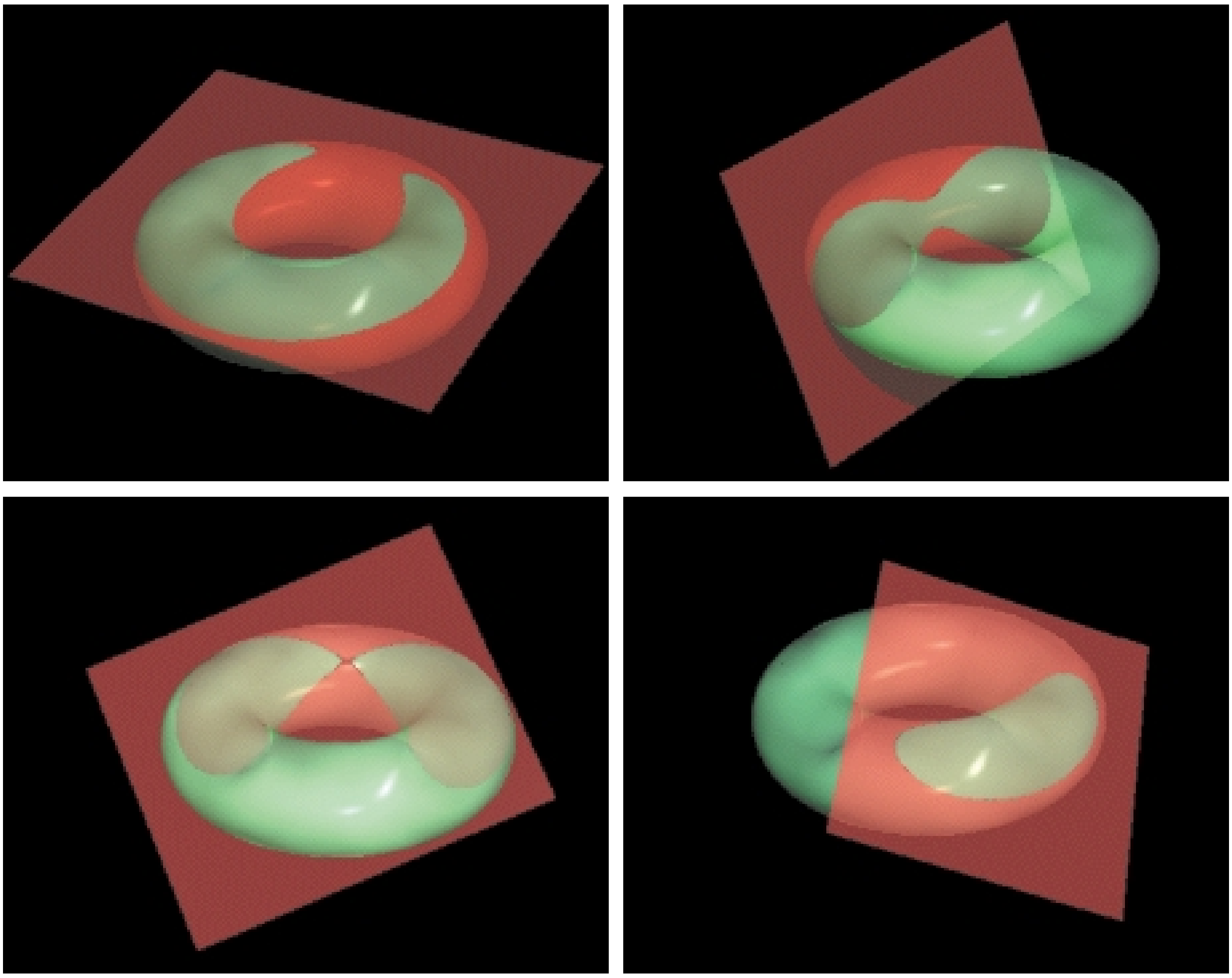


Figure 75. A few of the possible cuts through a bagel.

The importance of context is critical for the correct interpretation of data in scientific images. Our expectations based on prior experience and study, knowledge of how the sample was prepared and how the image was acquired, and even our hopes and fears about how an experiment may turn out, can significantly affect visual interpretation, and the recognition of features (or failure to recognize them).

Obviously, this works in two different ways. It is important to know enough about the sample and image to correctly interpret it, while avoiding the pitfalls that expectation can cause. Some people are better at this than others, and anyone can make the occasional mistake, but fortunately the open nature of scientific publication provides a mechanism for correction.

One very frequently encountered problem of context for images arises in microscopy. Sections are typically cut through tissue with a microtome for examination in transmission, or surfaces of opaque materials are prepared by polishing for examination by reflected light (or the equivalent use of the transmission or scanning electron microscope). In all of these cases, the images are inherently two-dimensional, but the structure that they represent and sample is three dimensional. It is very difficult for most people to provide a proper context for understanding these images in terms of the three-dimensional structure.

Nothing in our evolutionary experience has prepared our vision system to handle section images. We expect to see the outsides of objects. Shapes are implicitly understood as silhouettes. In fact, it is even difficult to mentally construct the appearance of sections through objects. An experiment that I have used with many classes of students is to place a bagel (a torus) in front of them and ask them to sketch about a dozen random sections through it (Figure 75). Everyone gets the standard "bagel cut" right, and after only a moment's thought most draw the two circular sections produced by a vertical cut. But then things get hard. There is a strong tendency to selectively draw sections that pass through the geometric center and are hence symmetrical. Few of the complex arcs and lopsided ovoids are represented. And many students draw sections that are not actually possible.

If it is that hard to imagine what the sections through a very simple shape will look like, how difficult is it for the complex shapes that occur in natural objects, such as organelles in cells or dendrites in metals? And this is the "forward" problem, going from a known three-dimensional shape to the two-dimensional sections that would result. It is much harder to go in reverse, seeing the two-dimensional shapes and having to imagine what the three-dimensional object must have been that was responsible for them.

In fact this is not always a unique solution. Something as simple as ellipses illustrates the situation. Elliptical sections will result from sections through either oblate or prolate ellipsoids of revolution (the prolate shape is like an American football, the oblate shape like a discus). If all of the ellipsoids are the same size and shape, then by examining a large number of random sections it is possible to determine whether the generating solid is oblate or prolate. If the length of the most elongated ellipse is similar in dimension to the diameter of the most equiaxed section, then the ellipsoid is oblate. If the width of the most elongated ellipse is close to the diameter of the equiaxed sections, the ellipsoid is prolate. This is intentionally not illustrated to force the reader to try to mentally construct these very simple shapes.

If the sizes of the ellipsoids vary, the problem becomes much more difficult. And if the shapes of the ellipsoids can also vary, it becomes insoluble. But regardless of the ability to intellectually solve a tricky three-dimensional puzzle, this is never a task that the human brain solves automatically based on experience with section images. All of our routine experience is with projected views of the outsides of objects, and it leads to misunderstandings about 3D structure examined by sectioning. As a simple example, generations of textbooks described mitochondria as "football-shaped" structures within the cell, because the most easily recognized sections through mitochondria appear as elliptical in shape. In fact, the 3D structure is cylindrical, with bends and branches that produce many types of sections, the most irregular of which are typically not recognized at all in sections.

It is possible to measure quite a few geometric properties of three-dimensional structure from two-dimensional images, including volumes, surface areas, curvature and length, the mean size of arbitrary and variable three-dimensional objects, and so on. The entire field of Stereology, several professional journals, and an international society, all exist to deal with this need. Methods of great power and in some cases surprising simplicity have been developed to accomplish the necessary measurements and calculations, and they are gradually becoming more widely used as microscopists realize the need for them. But they can never compensate for the fact that people just don't automatically understand the relationships between three-dimensional structure and two-dimensional images.

This seems to be true even after extensive training and developing familiarity with particular three dimensional structures. Medical doctors and technicians rely upon section images from instruments such as magnetic resonance imaging (MRI) and computed X-ray tomography (CAT scans) to study the human body. But it appears from tests and interviews that few of these people have a three-dimensional mental picture of the structure. Instead, they learn to recognize the normal appearance of sections that are almost always taken in a few standard orientations, and to spot deviations from the norm, particularly ones associated with common diseases or other problems.

Arrangements must be made

One thing that people are extremely good at, however, is finding order in an arrangement of objects. Sometimes this quest for simplification finds true and meaningful relationships between objects, and sometimes it does not. Historically, the construction of constellations by playing connect-the-dots among the bright stars in the sky seems to have been carried out by many different cultures (of course, with different results). Figure 76 shows the classical Greek version. Assembling the data needed to construct Stonehenge as a predictor of solstices and eclipses must have taken multiple lifetimes. The Copernican revolution that allowed ellipses to simplify the increasingly complex Ptolemaic circles-and-epicircles model of planetary motion was a quest for this same type of simplification.

Most scientists follow Einstein's dictum that it is important to find the simplest solution that works, but not one that is too simple. The idea is much older than that. William of Occam's "principle of parsimony" is that "one should not increase, beyond what is

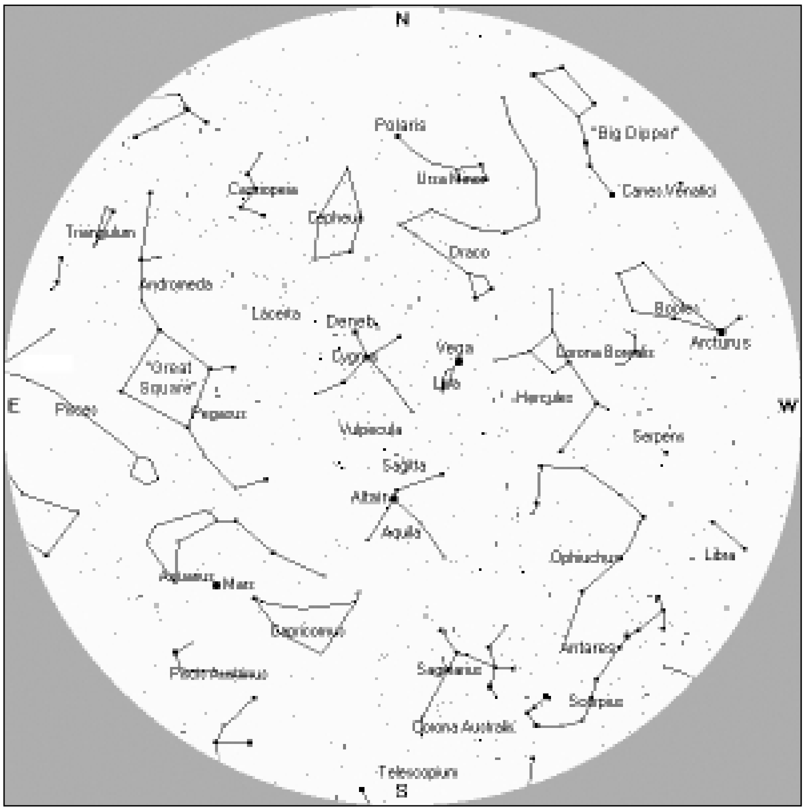


Figure 76. The familiar stellar constellations.

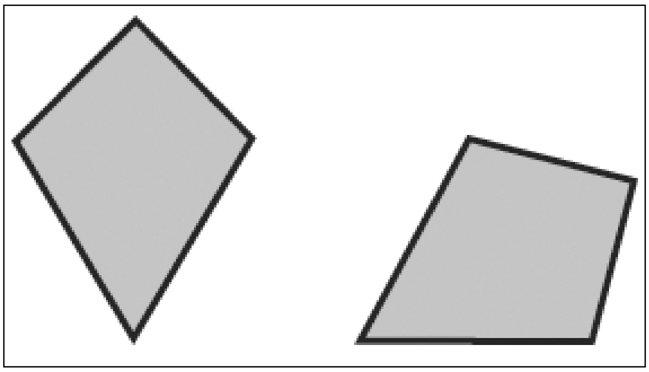


Figure 77. These two shapes are not perceived to be identical. The "kite shape" on the left has a dominant vertical axis of symmetry. The irregular four-sided polygon on the left has a horizontal base.

necessary, the number of entities required to explain anything." Instinctively we all seek the simple answer, sometimes in situations where there is not one to be found. And of course, this applies to the examination of images, also.

Finding visual alignments of points in images, or in plots of data, people prefer straight lines or smooth, gradual curves. Linear regression is probably the most widely used (and abused) method of data interpretation, for imposing order on a collection of points. Filling in gaps in lines or boundaries is often a useful procedure, but it can lead to mistakes as well. The illusory Kanisza triangles are an example of filling in gaps and connecting points. The process of connecting points and lines is closely related to grouping, which has been discussed before, for instance in the process of forming a number from the colored circles in the color blindness test images.

Human vision has a built-in directional bias that prefers the vertical, followed by the horizontal, and a strong preference for symmetry. As shown in Figure 77, that can bias our judgment. It also influences our ability to detect gradients or clustering.

Some processes produce a random distribution of features, like sprinkling salt onto a table. If every feature is completely independent of all the others, a random distribution results. Non-random distributions occur because features either attract or repel each other. Cacti growing in the desert are self-avoiding, because each one tries to protect its supply of water and nutrients. Particles floating on a liquid surface may tend to cluster because of surface tension effects. In extreme cases, visual observation of clustering or self-avoidance is possible. But people do not easily see through apparently chaotic distributions to detect these effects (Figure 78). In fact, the presence of variation in any type of feature appearance makes it very hard for visual inspection to detect an underlying order.

That isn't to say that people prefer the order. Tests with computer-generated images of apparently random paint droplets, showed that completely ordered images were considered boring and not visually stimulating, while completely random ones were considered to be equally uninteresting. When the correlation between size, color and position obeyed a "pink noise" or fractal relationship, the pictures were most visually interesting to viewers. Interestingly, the same relationships were found in mathematical analysis of several of Jackson Pollock's paintings (Figure 79).

Apparently a distribution in which there is just enough hint of order that we must work to find it visually is the most appealing. The implication for noticing (or failing to notice) arrangements of features in scientific images is clear; computer measurement can detect these properties, but only if we notice something that leads us to perform the necessary measurements.

Beyond the subject of arrangements that may be present throughout an image, gradients are often important, but not always easy to detect. The problem is that there can be so many different kinds of gradient.

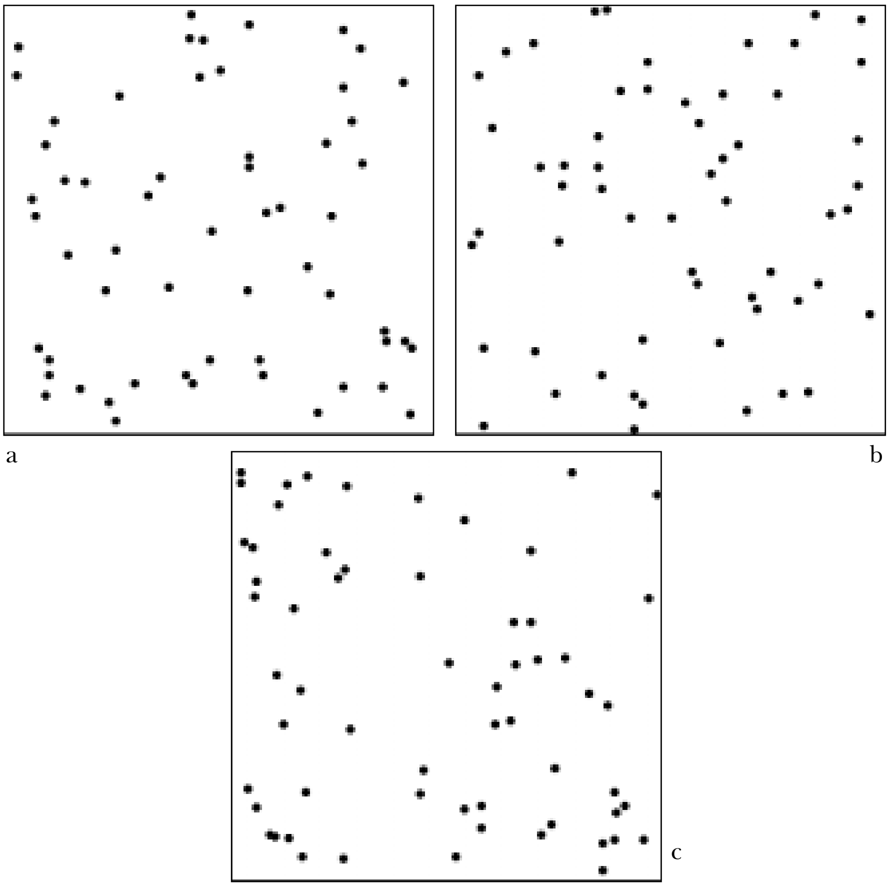


Figure 78. Examples of features distributions that are statistically a) clustered; b) self-avoiding; c) random

Figure 79. Jackson Pollock's "Blue Poles #11." His paintings have been analyzed to determine that they have a fractal structure and a complexity that evolved during his career.

Features may vary in size, shape, orientation, color, density, number, or any combination of these factors, as a function of position. Figure 80 shows only a few of the possibilities. The gradient may be linear, although not necessarily vertical or horizontal, but it can also be radial or follow a more complex path. In many cases, features have a tendency to cluster either near (or, conversely, away from) the boundary of an irregular region. Figure 81 illustrates the case of organelles in a cell, but the same thing occurs for plants near an aquifer, particles in metal grains, and people in a room at a party (depending on where the bar is located). Detecting the gradient in such cases is visually difficult and it is usually necessary to resort to measurement. Computer methods can determine the distance of each feature from a point or boundary, and this distance can then be used, along with the appropriate measures of size, shape, etc., to statistically assess whether a significant gradient is present.

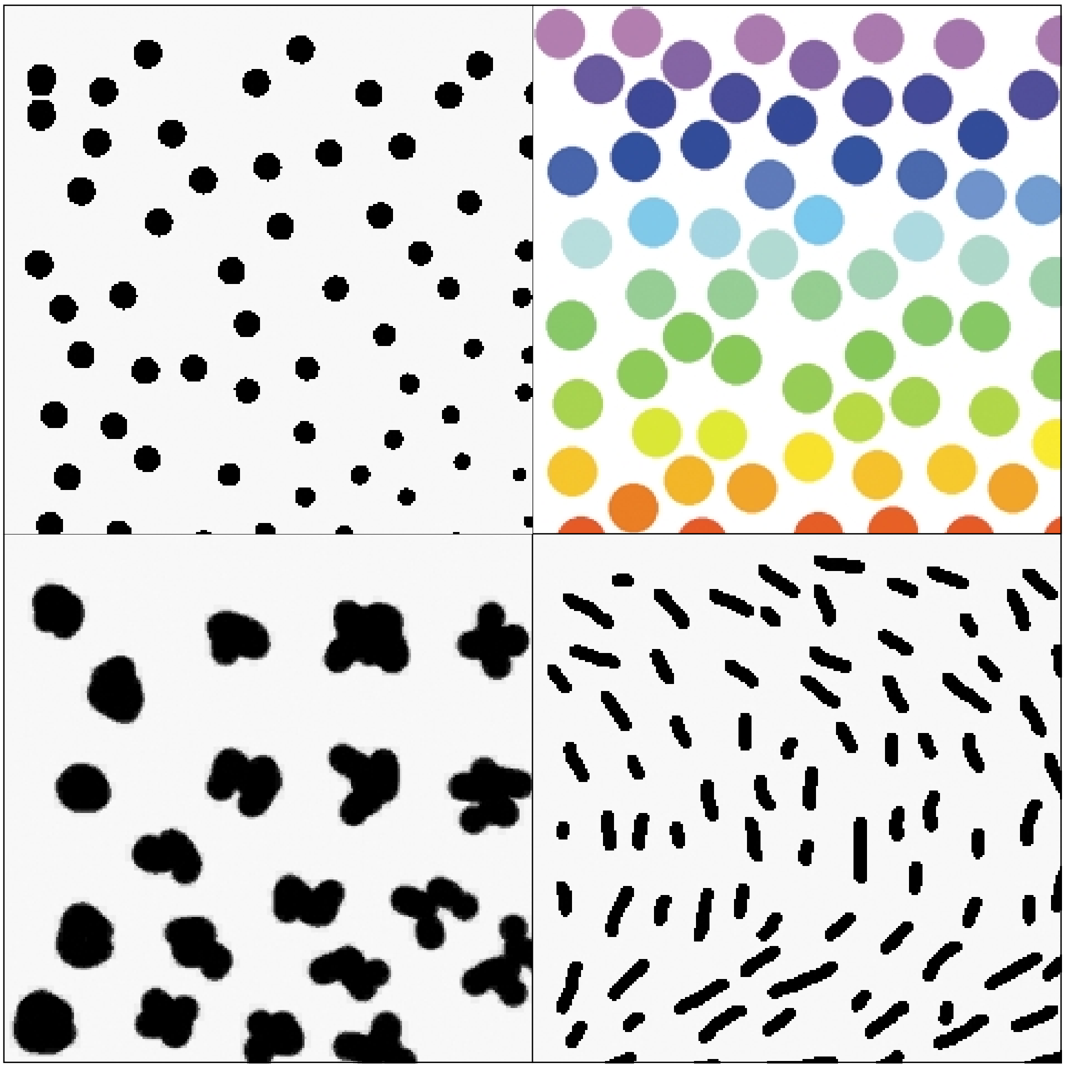


Figure 80. Examples of gradients of size, color, shape, and orientation.

The innate ability that people have for finding order in images is risky. Sometimes we imagine a regularity or order that is not there (e.g., constellations), and sometimes the presence of complexity or superficial disorder hides the real underlying structure. Some orientations are more readily detected than others, and complicated gradients are likely to escape detection unless we know beforehand what to

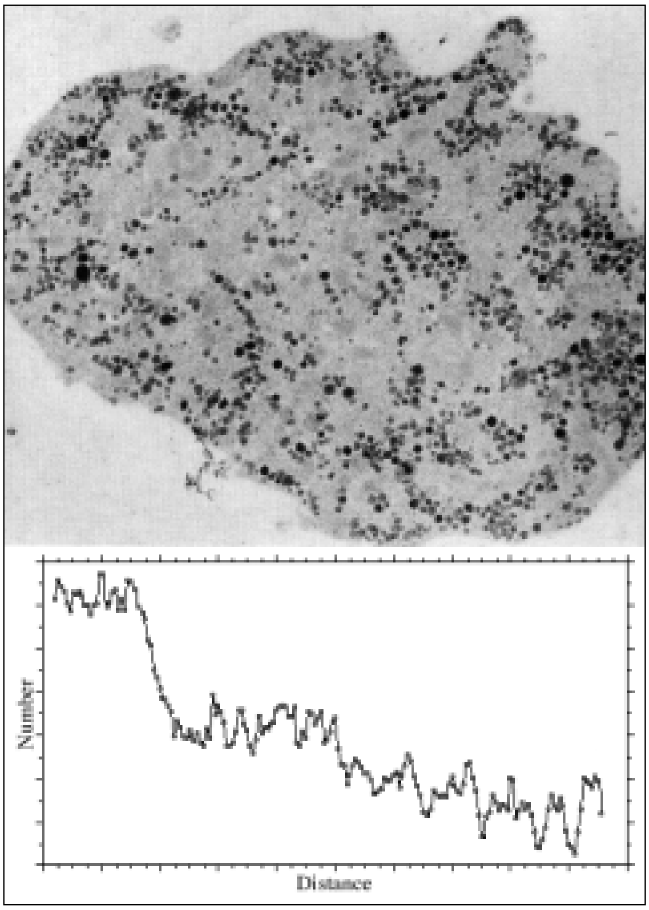


Figure 81. Distribution of organelles in a cell, with plot of number vs. distance.

look for. The result is that many real spatial arrangements may be missed, even in two dimensions (and because of the additional problems introduced by examining two-dimensional sections through three-dimensional structures, the problem is much worse for three-dimensional spatial arrangements).

So in conclusion...

Human vision is an extremely powerful tool, evolved over millenia to extract from scenes those details that are important to our survival as a species. The processing of visual information combines a hierarchy of highly parallel neural circuits to detect and correlate specific types of detail within images. Many short cuts that work "most of the time" are used to speed recognition. Studying the failure of these tricks, revealed in various visual illusions, aids in understanding of the underlying processes.

An awareness of the failures and biases is also important to the scientist who relies on visual examination of images to acquire or interpret data. Visual inspection is a comparative, not a quantitative process, and it is easily biased by the presence of other information in the image. Computer image analysis methods are available that overcome most of these specific problems, but they provide answers that are only as good as the questions that are asked. In most cases, if the scientist does not visually perceive the features or trends in the raw images, their subsequent measurement will not be undertaken.

For further reading

There are several classic books that provide a good introduction to human vision, without delving too deeply into the fascinating but specialized literature regarding messy anatomical details of the visual cortex. They include:

John P. Frisby (1980) Illusion, Brain and Mind, Oxford Univ. Press
Irvin Rock (1984) Perception, W. H. Freeman Co
David Marr (1982) Vision, W. H.Freeman Co.

Computer-based image analysis offers many valuable insights into human vision, if only to show that there are computational models that function differently but attempt with varying degrees of success to extract the same types of information. There are a great many books in this field. A good sampling with different topical coverage and styles would include

John C. Russ (2002) The Image Processing Handbook, 4th edition, CRC Press
Kenneth R. Castleman (1996) Digital Image Processing, Prentice Hall
Gaurav Sharma, ed. (2003) Digital Color Imaging Handbook, CRC Press
Alan Watt & Fabio Policarpo (1998) The Computer Image, Addison Wesley
Rafael C. Gonzalez & Richard E. Woods (1993) Digital Image Processing, Addison Wesley

Probably the most accessible texts covering various models for object recognition, and the software for implementing them, are:

Keinosuke Fukunaga (1990) Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press
Sing-Tze Bow (1992) Pattern Recognition and Image Processing, Marcel Dekker

To learn more about the science of stereology,the relationship between three-dimensional structure and two-dimensional images, see:

John C. Russ & Robert T. Dehoff (2002) Practical Stereology, 2nd edition, Plenum Press

Of course, there is also an extensive literature in many peer-reviewed journals, and in the modern era no one should neglect to perform a google search of the internet, which will locate several sets of course notes on this topic as well as publication reprints and many sites of varying quality.