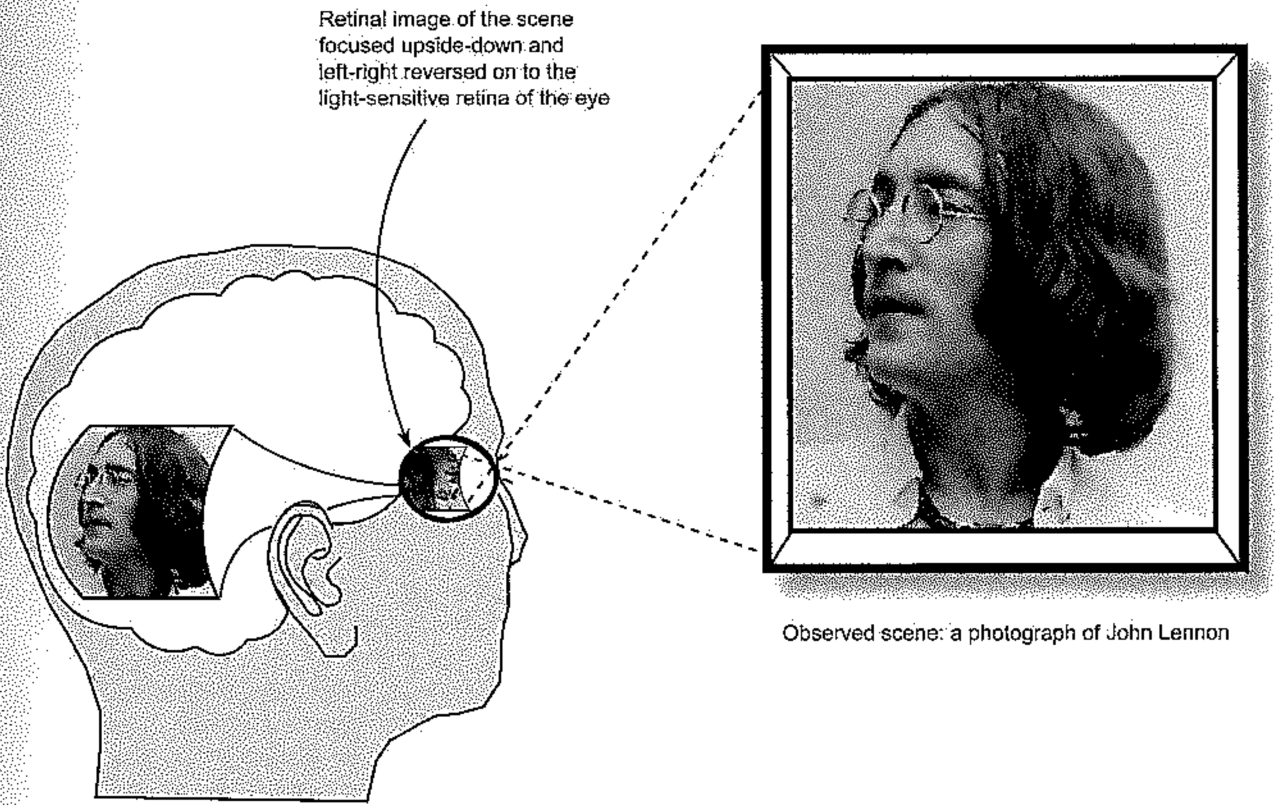
1

Seeing: What Is It?



Retinal image of the scene focused upside-down and left-right reversed on to the light-sensitive retina of the eye

Observed scene: a photograph of John Lennon

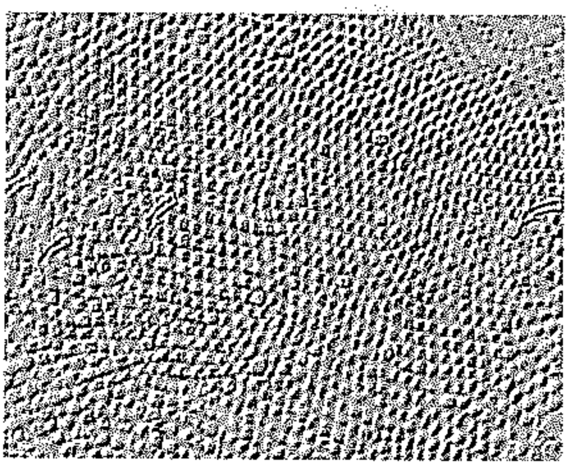
1.1 An "inner screen" theory of seeing
One theory of this kind proposes that there is a set of brain cells whose level of activity represents the brightness of points in the scene. This theory therefore suggests that seeing is akin to photography. Note that the image of Lennon is inverted in the eye, due to the optics of the eye, but it is shown upright in the brain to match our perceptions of the world—see page 8. Lennon photograph courtesy Associated Newspapers Archive.

What goes on inside our heads when we see? Most people take seeing so much for granted that few will ever have considered this question seriously. But if pressed to speculate, the ordinary person who is not an expert on the subject might suggest:

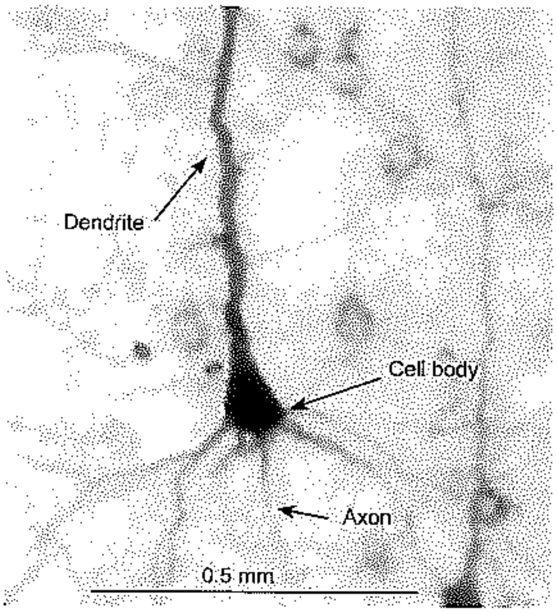
> Could perhaps there be an "inner screen" of some sort in our heads, rather like a cinema screen except that it is made out of brain tissue? The eyes transmit an image of the outside world onto this screen, and this is the image of which we are conscious?

The idea that seeing is akin to photography, illustrated in **1.1**, is commonplace, but it has fundamental shortcomings. We discuss them in this opening chapter and we introduce a very different concept about seeing.

The photographic metaphor for seeing has its foundation in the observation that our eyes are in many respects like cameras. Both camera and eye have a lens; and where the camera has a light-sensitive film or an array of light-sensitive electronic components, the eye has a light-sensitive retina, **1.2**, a network of tiny light-sensitive receptors arranged in a layer toward the back of the eyeball (Latin *rete*—net). The job of the lens is to focus an image of the outside world—the retinal image—on to these receptors. This image stimulates them so that each receptor encodes the intensity of the small point of light in the image that lands on it. Messages about these point by point intensities are conveyed from the eye along fibers in the optic nerve to the brain. The brain is composed of millions of tiny components, brain cells called ***neurons***, **1.3**.

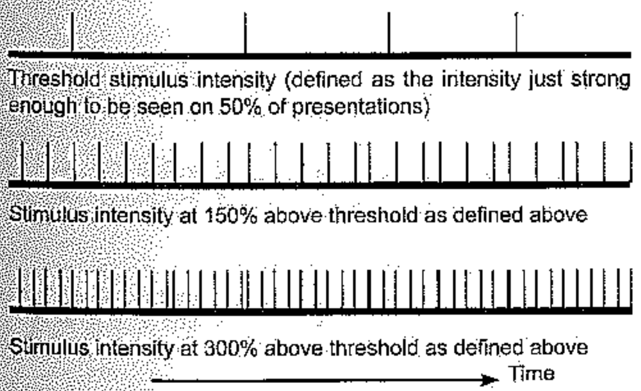


1.2 The receptor mozaic
Microphotograph of cells in the center of the human retina (the fovea) that deals with straight ahead vision. Magnification roughly x 1200. Courtesy Webvision (http://webvision.med.utah.edu/sretina.html#central).



1.3 Pyramidal brain cell
Microscopic enlargement of a slice of rat brain stained to show a large neuron called a ***pyramidal cell.*** The long thick fiber is a dendrite that collects messages from other cells. The axon is the output fiber. (*Note* Some other types of neurons have thicker axons.) Brain neurons are highly interconnected: it has been estimated that there are more connections in the human brain than there are stars in the Milky Way. Courtesy P. Redgrave.

The core idea of the "inner screen" theory illustrated in **1.1** is that certain brain cells specialize in vision and are arranged in the form of a sheet—the "inner screen." Each cell in the screen can at any moment be either active, inactive, or somewhere in between, **1.4**. If a cell is very active, it is signaling the presence of a bright spot at that particular point on the "inner screen"—and hence at the associated point in the outside world. Equally, if a cell is only moderately active, it is signaling an intermediate shade of gray. Completely inactive cells signal black spots. Cells in the "inner screen" as a whole take on a pattern of activity whose overall shape mirrors the shape of the retinal image received by the eye. For example, if a photograph is being observed, as in **1.1**, then the pattern of activ-


Threshold stimulus intensity (defined as the intensity just strong enough to be seen on 50% of presentations)
Stimulus intensity at 150% above threshold as defined above
Stimulus intensity at 300% above threshold as defined above
Time

1.4 Stimulus intensity and firing frequency
Most neurons send messages along their ***axons*** to other neurons. The messages are encoded in ***action potentials*** (each one is a brief pulse of voltage change across the neuron's outer "skin" or membrane). In the schematic recordings above, the time scale left-to-right is set so that the action potentials show up as single vertical lines or ***spikes***. The height of the spikes remains constant: this is the ***all-or-none law***—the spike is either present or absent. On the other hand, the ***frequency of firing*** can vary, which allows the neuron to use an ***activity code***, here for representing stimulus intensity. Firing rates for brain cells can vary from 0 to several hundred spikes per second.

ity on the "inner screen" resembles the photograph. The "inner screen" theory proposes that as soon as this pattern is set up on the screen of cells, the observer has the experience of seeing.

This "inner screen" theory of seeing is easy to understand and is intuitively appealing. After all, our visual experiences do seem to "match" the outside world: so it is natural to suppose that there are mechanisms for vision in the brain which provide the simplest possible type of match—a physically similar or "photographic" one. Indeed, the "inner screen" theory of seeing can also be likened to television, an image-transmission system which is also photographic in this sense. The eyes are equivalent to TV cameras, and the image finally appearing on a TV screen connected to the cameras is roughly equivalent to the proposed image on the "inner screen" of which we are conscious. The only important difference is that whereas the TV-screen image is composed of more or less brightly glowing dots, our visual image is composed of more or less active brain cells.

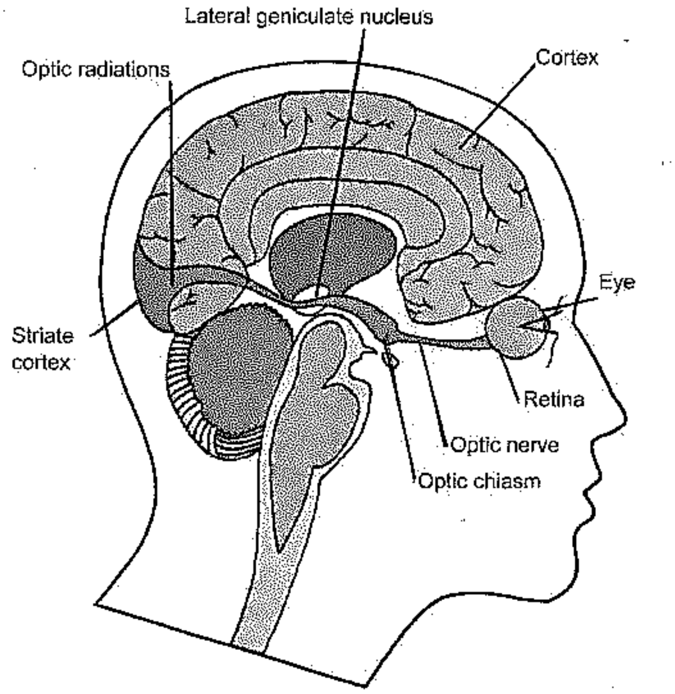
Seeing and Scene Representations

The first thing to be said about the "inner screen" theory of seeing is that it proposes a certain kind of ***representation*** as the basis of seeing. In this respect it is like almost all other theories of seeing, but to describe it in this way requires some explanation.

In this book we use the term representation for anything that stands for something other than itself. Words are representations in this sense. For example, the word "chair" stands for a particular kind of sitting support—the word is not the support itself. Many other kinds of representations exist, of course, apart from words. A red traffic light stands for the command "Stop!", the Stars and Stripes stands for the United States of America, and so on. A moment's reflection shows that there must be representations inside our heads which are unlike the things they represent. The world is "out there," whereas the perceptual world is the result of processes going on inside the pink blancmange-like mass of brain cells that is our brain. It is an inescapable conclusion that there must be a representation of the outside world in the brain. This representation can be said to serve as a *description* that encodes the various aspects of the world of which sight makes us aware.

In fact, when we began by asking "What goes on inside our heads when we see?" we could as well have stated this question as: "When we see, what is the nature of the representation inside our heads that stands for things in the outside world?" The answer given by the "inner screen" theory is that each brain cell in the hypothetical screen is describing (representing) the brightness of one particular spot in the world in terms of an ***activity code*, 1.4**. The code is a simple one: the more active the cell, the lighter or more brightly illuminated the point in the world.

It can come as something of a shock to realize that somehow the whole of our perceived visual world is tucked away in our skulls as an inner representation which stands for the outside world. It is difficult and unnatural to disentangle the "perception of a scene" from the "scene itself." Nevertheless, they must be clearly distinguished if seeing is to be understood. When the difference between a perception and the thing perceived is fully grasped, the conclusion that seeing must involve a representation of the viewed scene sitting somewhere inside our heads becomes easier to accept. Moreover, the problem of seeing can be clearly stated: what is the nature of the brain's representation of the visual



1.5 Diagrammatic section through the head
This shows principal features of the major visual pathway that links the eyes to the cortex.

world, and how is it obtained? It is this problem which provides the subject of this book.

Perception, Consciousness, and Brain Cells

One reason why it might feel strange to regard visual experience as being encoded in brain cells is that they may seem quite insufficient for the task. The "inner screen" theory posits a direct relationship between conscious visual experiences and activity in certain brain cells. That is, activity in certain cells is somehow accompanied by conscious experience. Proposing this kind of parallelism between brain-cell activity and visual experience is characteristic of many theories of perceptual brain mechanisms. But is there more to it than this? Can the richness of visual experience really be identified with activity in a few million, or even a few trillion, brain cells? Are brain cells the right kind of entities to provide conscious perceptual experience? We return to these questions in Ch 22. For the moment, we simply note that most vision scientists get on with the job of studying seeing without concerning themselves much with the issue of consciousness.

Pictures in the Brain

You might reasonably ask at this point: does neuroscience have anything to say directly about the "inner screen" theory? Is there any evidence from studies of the brain as to whether such a screen or anything like it exists?

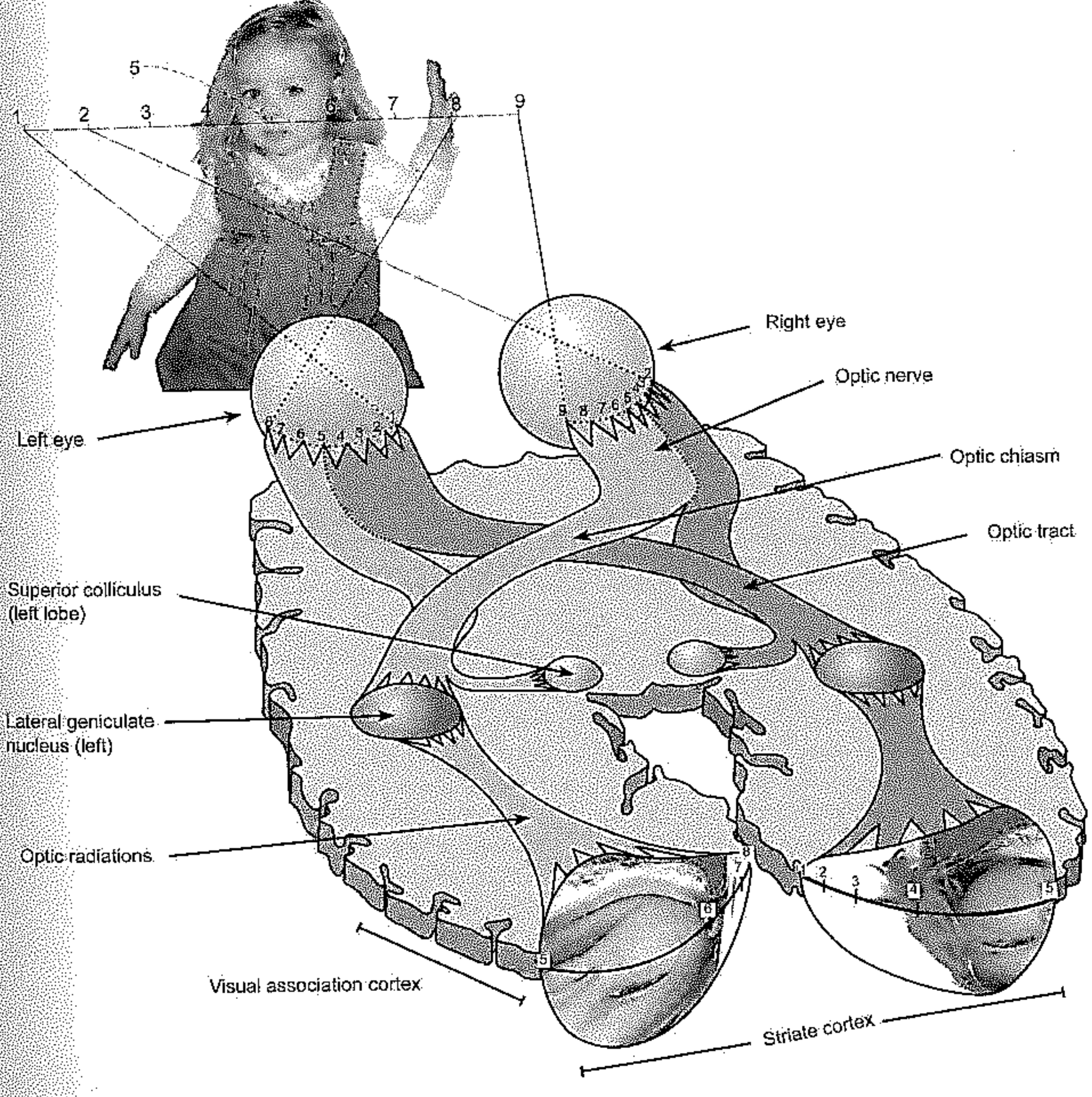
The major visual pathway carrying the messages from the eyes to the brain is shown in broad outline in **1.5**. Fuller details are shown in **1.6** in which the eyes are shown inspecting a person, and the locations of the various parts of this scene "in" the visual system are shown with the help of numbers.

The first thing to notice is that the eyes do not receive identical images. The left eye sees rather more of the scene to the left of the central line of sight (regions 1 and 2), and vice versa for the right eye (regions 8 and 9). There are other differences between the left and right eyes' images in the case of 3D scenes: these are described fully in Ch 18.

Next, notice the ***optic nerves*** leaving the eyes. The fibers within each optic nerve are the ***axons*** of certain retinal cells, and they carry messages from the retina to the brain. The left and right optic nerves meet at the ***optic chiasm***, **1.6** and **9.9**, where the optic nerve bundle from each eye splits in two. Half of the fibers from each eye cross to the *opposite* side of the brain, whereas the other half stay on the same side of the brain throughout.

The net result of this partial crossing-over of fibers (technically called ***partial decussation)*** is that messages dealing with any given region of the field of view arrive at a common destination in the cortex, regardless of which eye they come from. In other words, left- and right-eye views of any given feature of a scene are analyzed in the same physical location in the ***striate cortex***. This is the major receiving area in the cortex for messages sent along nerve fibers in the optic nerves.

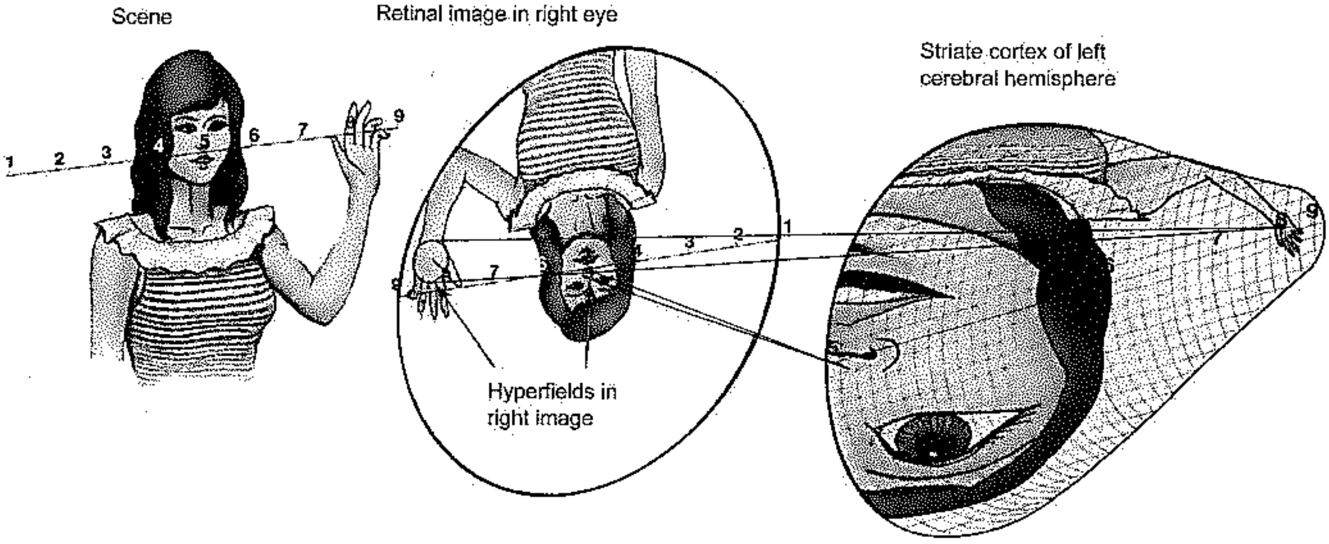
Fibers from the optic chiasm enter the left and right ***lateral geniculate nuclei***. These are the first "relay stations" of the fibers from the eyes on their way to the striate cortex. That is, axons from the retina terminate here on the dendrites of new neurons, and it is the axons of the latter cells that then proceed to the cortex. A good deal of mystery still



1.6 Schematic illustration of two important visual pathways
One pathway goes from the eyes to the striate cortex and one from the eyes to each superior colliculus. The distortion in the brain mapping in the striate cortex reflects the emphasis given to analysing the central region of the retinal image, so much so that the tiny representation of the child's hand can hardly be seen in this figure. See **1.7** for details.

surrounds the question of what cells in the lateral geniculate nuclei do. They receive inputs not only from the eyes but also from other sense organs, so some think that they might be involved in filtering messages from the eyes according to what is happening in other senses. The lateral geniculate nuclei also receive a lot of fibers sending messages from the cortex. Hence there is an intriguing two-way up-down traffic going on in this visual pathway and we discuss its possible functions in later chapters.

Before we go on to discuss the way fiber terminations are laid out in the striate cortex, note that the optic nerves provide visual information to two other structures shown in **1.6**—the left and right halves of the ***superior colliculus***. This is a


Scene
Retinal image in right eye
Striate cortex of left cerebral hemisphere
1
2
3
4
5
6
7
8
9
Hyperfields in right image

1.7 Mapping of the retinal image in the striate cortex (schematic)
Turn the book upside-down for a better appreciation of the distortion of the scene in cortex. The ***hyperfields*** are regions of the retinal image that project to hypothetical structures called ***hypercolumns*** (denoted as graph-paper squares in the part of the striate cortex map shown here, which derives from the left hand sides of the left and right retinal images; more details in Chs 9 and 10). Hyperfields are much smaller in central than in peripheral vision, so relatively more cells are devoted to central vision. Hyperfields have receptive fields in both images but here two are shown for the right image only.

brain structure which lies underneath the ***cortex,*** **1.5**, so it is said to be ***sub-cortical.*** Its function is different from that performed by regions of the cortex devoted to vision. The weight of evidence at present suggests that the superior colliculus is concerned with guiding visual attention. For example, if an object suddenly appears in the field of view, mechanisms within the superior colliculus detect its presence, work out its location, and then guide eye movements so that the novel object can be observed directly.

It is important to realize that other visual pathways exist apart from the two main ones shown in **1.6.** In fact, in monkeys and most probably also in man, optic nerve fibers directly feed at least six different brain sites. This is testimony to the enormously important role of vision for ourselves and similar species. Indeed, it has been estimated that roughly 60% of the brain is involved in vision in one way or another.

Returning now to the issue of pictures-in-the-brain, the striking thing in **1.6** is the orderly, albeit curious, layout of fiber terminations in the striate cortex.

First, note that a face is shown mapped out on the cortical surface (*cortical* means "of the cortex"). This is the face that the eyes are inspecting.

Second, the representation is upside-down. The retinal images (not shown in **1.6**) are also upside-down due to the way the optics of the eyes work, **1.1.** Notice that the sketch of the "inner screen" in **1.1** showed a right-way-up image, so it is different in that respect from the mapping found in the striate cortex.

Third, the mapping is such that the representation of the scene is cut right down the middle, and each half of the cortex (technically, each ***cerebral hemisphere***) deals initially with just one half of the scene.

Fourth, and perhaps most oddly, the cut in the representation places adjacent regions of the central part of the scene farthest apart in the brain!

Fifth, the mapping is spatially distorted in that a greater area of cortex is devoted to central vision than to peripheral: hence the relatively swollen face and the diminutive arm and hand, **1.7.** This doesn't mean of course that we actually *see* people in this distorted way—obviously we don't. But it reveals that a much larger area in our brain is assigned to ***foveal vision*** (i.e., analyzing what we are directly looking at) than is devoted to ***peripheral vision.*** This dedication of most cortical tissue to foveal vision is why we are much better at seeing details in the region of the scene we are looking at than we are at seeing details which fall out toward the edges of our field of view.

All in all, the cortical mapping of incoming visual fibers is curious but orderly. That is, adjacent regions of cortex deal with adjacent regions of the scene (with the exception of the mid-line cut). The technical term for this sort of mapping is ***topographic.*** In this instance it is called ***retinotopic*** as the mapping preserves the neighborhood relationships that exist between cells in the retina (except for the split down the middle). The general orderliness of the mapping is reminiscent of the "inner screen" proposed in **1.1.** But the oddities of the mapping should give any "inner screen" theorist pause for thought. The first "screen," if such it is, we meet in the brain is a very strange one indeed.

The striate cortex is not the only region of cortex to be concerned with vision—far from it. Fibers travel from the striate cortex to adjacent regions, called the ***pre-striate cortex*** because they lie just in front of the striate region. These fibers preserve the orderliness of the mapping found in the striate region. There are in fact topographically organized visual regions in the pre-striate zone and we describe these ***maps*** in Ch 10. For the present, we just note that each one seems to be specialized for a particular kind of visual analysis, such as color, motion, etc. One big mystery is how the visual world can appear to us as such a well-integrated whole if its analysis is actually conducted at very many different sites, each one serving a different analytic function.

To summarize this section, brain maps exist which bear some resemblance to the kind of "inner screen" idea hesitantly advanced by our fictional "ordinary person" who was pressed to hazard a guess at what goes on the brain when we see. However, the map shown in **1.6-1.7** is not much like the one envisaged in **1.1**, being both distorted, upside-down and cut into two.

These oddities seriously undermine the photographic metaphor for seeing. But it is timely to change now from looking inside the brain for an "inner screen" and to examine in detail serious logical problems with the "inner screen" idea as a theory of seeing. We begin this task by considering man-made systems for seeing.

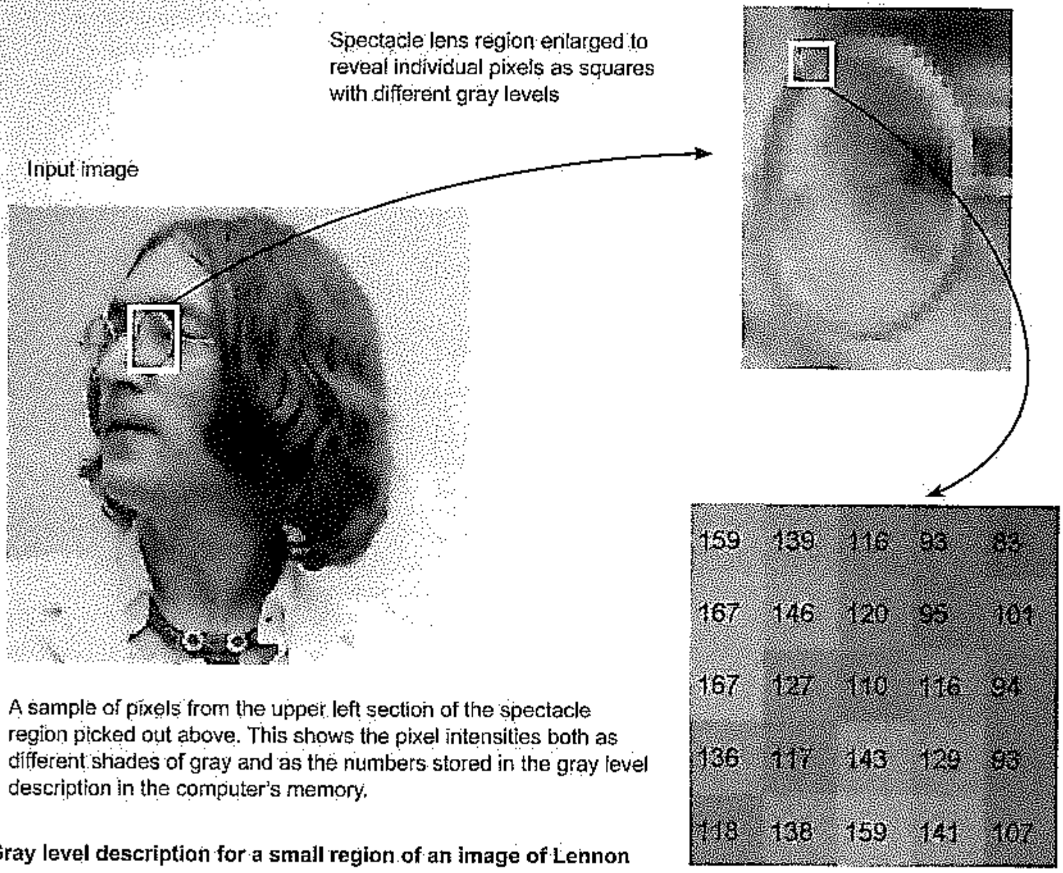
Machines for Seeing

A great deal of research has been done on building ***computer vision*** systems that can do visual tasks. These take in images of a scene as input, analyze the visual information in these images, and then use that information for some purpose or other, such as guiding a robot hand or stating what objects the scene contains and where they are. In our terminology, a machine of this type is deriving a scene description from input images.

Whether one should call such a device a "perceiver," a "seeing machine," or more humbly an "image processor" or "pattern recognizer," is a moot point which may hinge on whether consciousness can ever be associated with non-biological brains. In any event, scientists who work on the problem of devising automatic image-processing machines would call the activity appearing on the "inner screen" of **1.1** a kind of ***gray level description*** of the painting. This is because the "inner screen" is a representation signaling the various shades of gray all over the picture, **1.8.** (We ignore color in the present discussion, and also many intricacies in the perception of gray: see Ch 16). Each individual brain cell in the screen is describing the gray level at one particular point of the picture in terms of an ***activity code.*** The code is simple: the lighter or more brightly illuminated the point in the painting, the more active the cell.

The familiar desktop image scanner is an example of a human-made device that delivers gray level descriptions. Its optical sensor sweeps over the image laid face down on its glass surface, thereby measuring gray levels directly rather than from a lens-focused image. Their scanning is technically described as a ***serial*** operation as it deals with different regions of the image in sequence.

Digital cameras measure the point by point intensities of images focused on their light sensitive surfaces, so in this regard they are similar to biological eyes. They are said to operate in ***parallel*** because they take their intensity measurements everywhere over the image at the same instant. Hence they can deliver their gray levels quickly.


Spectacle lens region enlarged to reveal individual pixels as squares with different gray levels
Input image

159	139	116	93	83
167	146	120	95	101
167	127	110	116	94
136	117	143	129	93
118	138	159	141	107

A sample of pixels from the upper left section of the spectacle region picked out above. This shows the pixel intensities both as different shades of gray and as the numbers stored in the gray level description in the computer's memory.

1.8 Gray level description for a small region of an image of Lennon

The intensity measurements taken by both scanners and digital cameras are recorded as numbers stored in a digital memory. To call this collection of numbers a "gray level description" is apt because this is exactly what the numbers are providing, as in **1.8**.

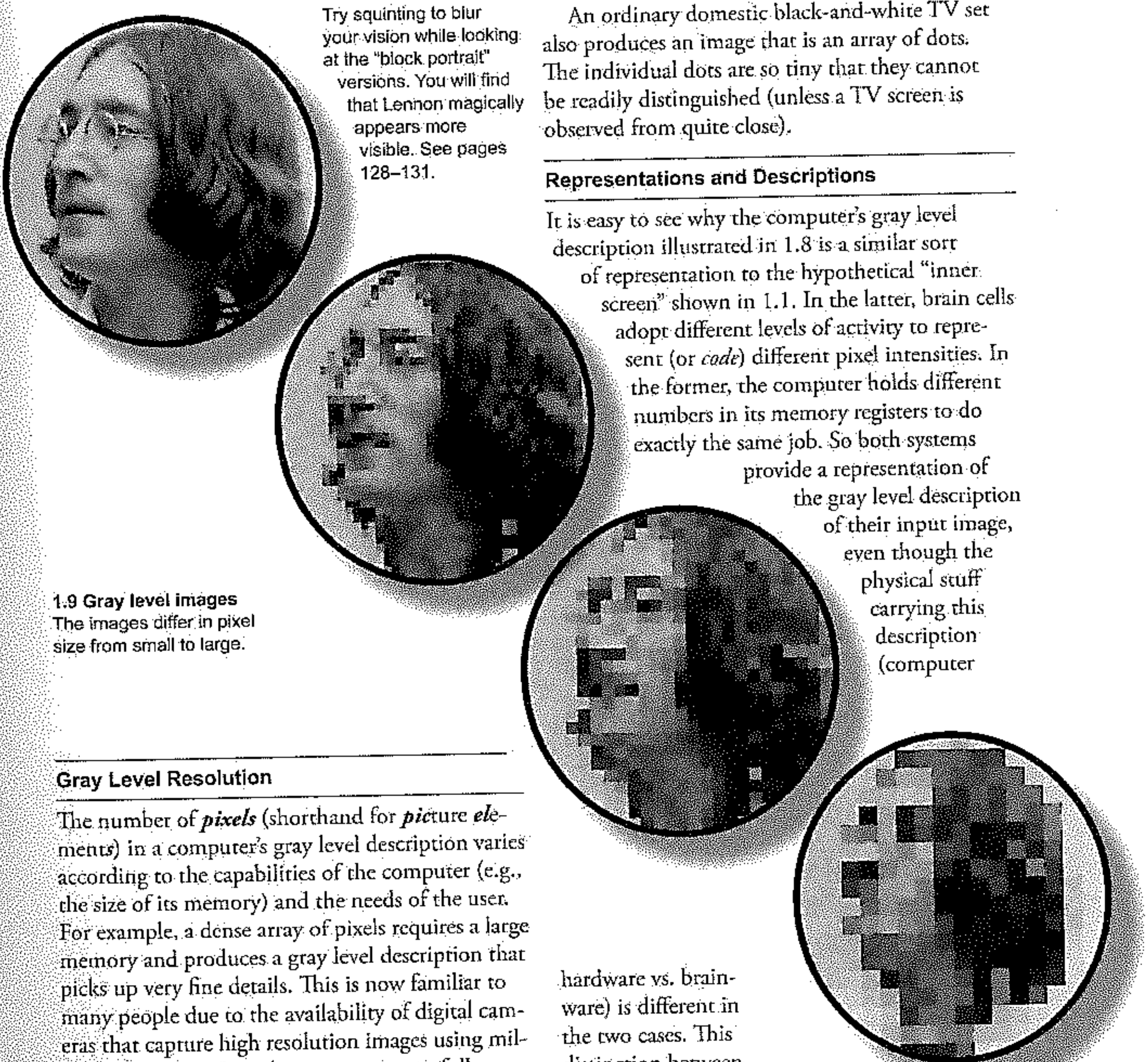
The term "gray level" arises from the black-and-white nature of the system, with black being regarded as a very dark gray (and recorded with a small number) and white as a very light gray (and recorded with a large number).

The numbers are a description in the sense defined earlier: they make *explicit* the gray levels in the input image. That is, they make these gray levels immediately usable (which means there is no need for further processing to recover them) by subsequent stages of image processing.

Retinal images are upside-down, due to the optics of the eyes (Ch 2) and many people are worried by this. "Why doesn't the world therefore appear upside down?", they ask.

The answer is simple: as long as there is a systematic correspondence between the outside scene and the retinal image, the processes of image interpretation can rely on this correspondence, and build up the required scene description accordingly. Upside-down in the image is simply interpreted as right-way-up in the world, and that's all there is to it.

If an observer is equipped with special spectacles which optically invert the retinal images so that they become the "right-way-up," then the world appears upside-down until the observer learns to cope with the new correspondence between image and scene. This adjustment process can take weeks, but it is possible. The exact nature of the adjustment process is not yet clear: does the upside-down world really begin to "look" right-way-up again, or is it simply that the observer learns new patterns of adjusted movement to cope with the strange new perceptual world he finds himself in?



Try squinting to blur your vision while looking at the "block portrait" versions. You will find that Lennon magically appears more visible. See pages 128–131.

1.9 Gray level images The images differ in pixel size from small to large.

Gray Level Resolution

The number of ***pixels*** (shorthand for ***pic****ture* ***el****ements*) in a computer's gray level description varies according to the capabilities of the computer (e.g., the size of its memory) and the needs of the user. For example, a dense array of pixels requires a large memory and produces a gray level description that picks up very fine details. This is now familiar to many people due to the availability of digital cameras that capture high resolution images using millions of pixels. When these are output as full-tone printouts, the images are difficult to discriminate from film-based photographs.

If fewer pixels are used, so that each pixel represents the average intensity over quite a large area of the input image, then a full-tone printout of the same size takes on a block-like appearance. That is, these images are said to show ***quantization*** effects. These possibilities are illustrated in **1.9**, where the same input image is represented by four different gray level images, with pixel arrays ranging from high to low resolution.

An ordinary domestic black-and-white TV set also produces an image that is an array of dots. The individual dots are so tiny that they cannot be readily distinguished (unless a TV screen is observed from quite close).

Representations and Descriptions

It is easy to see why the computer's gray level description illustrated in 1.8 is a similar sort of representation to the hypothetical "inner screen" shown in 1.1. In the latter, brain cells adopt different levels of activity to represent (or *code*) different pixel intensities. In the former, the computer holds different numbers in its memory registers to do exactly the same job. So both systems provide a representation of the gray level description of their input image, even though the physical stuff carrying this description (computer hardware vs. brainware) is different in the two cases. This distinction between the functional or design status of a representation (the job it performs) and the physical embodiment of the representation (different in man or machine of course) is an extremely important one which deserves further elaboration.

Consider, for example, the physical layout of the hypothetical "inner screen" of brain cells. This is an anatomically neat one, with the various pixel cells arranged in a format which physically matches the arrangement of the corresponding image points.

In sharp contrast to this, the computer registers that perform the same job as the hypothetical brain cells would not be arranged in the computer in a way which physically matches the input image. That is, the "anatomical" locations of the registers in the computer memory would not necessarily be arranged as the hypothetical brain cells are, in a grid-like topographical form that preserves the neighbor-to-neighbor spatial relationships of the image points.

Instead, the computer registers might be arranged in a variety of different ways, depending on many different factors, some of them to do with how the memory was manufactured, others stemming from the way the computer was programmed to store information. The computer keeps track of each pixel measurement in a very precise manner by using a system of labels (technically, *pointers*) for each of its registers, to show which part of the image each one encodes. The details of how this is done do not concern us: it is sufficient to note that the labels ensure that each pixel value can be retrieved for later processing as and when required. Consequently, it is true to say that the hypothetical brain cells of **1.1** and the receptors of **1.2** are serving the same *representational function* as the computer memory registers of **1.8**—recording the gray level of each pixel—even though the *physical nature* of the representation in each case differs radically. It differs in both the nature of the pixel code (cell activity versus size of stored number) and the anatomical arrangement of the entities that represent the pixels.

The idea that different physical entities can mediate the same information processing tasks is the fundamental assumption underlying the field of ***artificial intelligence***, which can be defined as the enterprise of making computers do things which would be described as intelligent if done by humans.

Before leaving this topic we note another major difference between the putative brain cells coding the gray level description and computer memory registers. Computers are built with an extremely precise organization of their components. As stated above, each memory register has a label and its contents can be set to represent different things according to the program being run on the computer. One moment the register might be holding a number within a spreadsheet, a few moments later it might be holding the code for a letter in document being edited using a word processor, or whatever. Indeed, the capacity for the *arbitrary* assignment of computer registers, to hold different contents that mean different things at different times according to the particular computation being run, is held by some to be the true hallmark of ***symbolic computation***.

But this capacity for arbitary and changing assignment is quite unlike the brain cells supporting vision, which, as far as we presently understand things, are more or less permanently committed to serve a *particular* visual function (but see the caveat below on learning). That is, if a brain cell is used to represent a scene property, such as the orientation of an edge, then that is the job that cell always does. It isn't quickly reassigned to represent, say, a dog, or a sound, etc., under the control of other brain processes.

It may be that other brain regions do contain cells whose functional role changes from moment to moment (perhaps cells supporting language?). If so, they would satisfy the arbitrary assignment definition of a symbolic computational device given above. However, some have doubted whether the brain's wiring really can support the highly accurate cell-to-cell connections that this would require. In any event, visual neurons do not appear to have this property and we will not use this definition of symbolic computation in this book.

A caveat that needs to be posted here is to do with various phenomena in ***perceptual learning***: we get better at various visual tasks as we practice them, and this must reflect changes in vision brain cells. Also, ***plasticity*** exists in brain cell circuits in early development, Ch 4, and parts of the brain to do with vision may even be taken over for other functions following blindness caused by losing the eyes (or vice versa: the visual brain may encroach on other brain areas).

But this caveat is about slowly acting forms of learning and plasticity. It does not alter the basic point being made here. When we say the visual brain is a symbolic processor we are *not* saying that its brain cells serve as symbols in the way that programmable computer components serve symbolic functions using different symbolic assignments from moment to moment.

Levels of Understanding Complex Information Processing Systems

Why have we dwelt on the point that certain brain cells and computer memory registers could serve the same task (in this case representing point-by-point image intensities) despite huge differences in their physical characteristics?

The answer is that it is a good way of introducing the linking theme running through this book. This is that we need to keep clearly distinct *different levels of discourse* when trying to understand vision.

This general point has had many advocates. For example, Richard Gregory, a distinguished scientist well known for his work on vision, pointed out long ago that understanding a device such as a radio needs an understanding of its individual components (transistors, resistors, etc.) and also an understanding of the design used to connect these components so that they work as a radio.

This point may seem self-evident to many readers but when it comes to studying the brain some scientists in practice neglect it, believing that the explanation must lie in the "brainware." Obviously, we need to study brain structures to understand the brain. But equally, we cannot be said to understand the brain unless we understand, among other things, *the principles underlying its design*.

Theories of the design principles underlying seeing system are often called ***computational theories***. This term fits analyzing seeing as an information processing task, for which the inputs are the images captured by the eyes and the outputs are various representations of the scene.

Often it is useful to have a level of analysis of seeing intermediate between the computational theory level and the hardware level. This level is concerned with devising good procedures or ***algorithms*** for implementing the design specified by the computational theory. We will delay specifying what this level tries to do until we give specific examples in later chapters.

What each level of analysis tries to achieve will become clear from the numerous examples in this book. We hope that by the time you have finished reading it we will have convinced you of the importance of the computational theory level for understanding vision. Moreover, we hope we will have given a number of sufficiently well-worked out examples to illustrate its importance when it comes to understanding vision. For the moment, we leave this issue with a famous quotation from an influential computational theorist, David Marr, whose approach to studying vision provides the linking theme for this book:

> Trying to understand vision by studying only neurons is like trying to understand bird flight by studying only feathers: it just cannot be done. (Marr, 1982)

Representing Objects

The "inner screen" of **1.1** can, then, be described as a particular kind of symbolic scene representation. The activities of the cells which compose the "screen" describe in a symbolic form the intensities of corresponding points in the retinal image of the scene being viewed. Hence, the theory proposes, these cell activities represent the lightnesses of the corresponding points in the scene. We are now in a position to see one reason why this is such an inadequate theory of seeing: it gives us such a woefully impoverished scene description!

The scene description which exists inside our heads is *not* confined simply to the lightnesses of individual points in the scene before us. It tells us an enormous amount more than this. Leaving aside the already noted limitation of not having anything to say about color, the "inner screen" description does not help us understand how we know what *objects* we are looking at, or how we are able to describe their various features—shape, texture, movement, size—or their spatial relationships one to another. Such abilities are basic to seeing—they are what we have a visual system for, so that sight can guide our actions and thoughts. Yet the "inner screen" theory leaves them out altogether.

You might feel tempted to reply at this point: "I don't really understand the need to propose anything more than an "inner screen" in order to explain seeing. Surely, once this kind of symbolic description has been built up, isn't that enough? Are not all the other things you mention—recognizing objects and so forth—an immediately given consequence of having the photographic type of representation provided by the "inner screen"?"

One reply to this question is that the visual system is so good at telling us what is in the world around us that we are understandably misled into taking its effortless scene descriptions for granted. Perhaps it is because vision is so effortless for us that is tempting to suppose that the scene we are looking at is "immediately given" by a photographic type of representation. But the truth is the exact opposite. Arriving at a scene description as good as that provided by the visual system is an immensely complicated process requiring a great deal of interpretation of the often limited information contained in gray level images. This will become clear as we proceed through the book. Achieving a gray level description of images formed in our eyes is only the first and easiest task. It is served by the very first stage of the visual pathway, the light-sensitive receptors in the eyes, **1.2**.

This point is so important it is worth reiterating. The intuitive appeal of the "inner screen" theory lies in its proposal that the visual system builds up a photographic-type of brain picture of the observed scene, and its suggestion that this brain picture is the basis of our conscious visual experience. The main trouble with the theory is that although it proposes a symbolic basis for vision, the symbols it offers correspond to points in the scene. Everything else in the scene is left unanalyzed and not represented *explicitly*.

It is not much good having the visual system build a photographic-type copy of the scene if, when that task is done, the system is no nearer to using information in the retinal image to decide what is present in the scene and to act appropriately, e.g., avoiding obstacles or picking up objects. After all, we started the business of seeing in the retina with a kind of picture, the activity in the receptor mosaic. It doesn't take us any further toward using vision to guide action to propose a brain picture more or less mirroring the retinal one. The inner screen theory is thus vulnerable to what philosophers call an ***infinite regress***: the problems with the theory cannot be solved by positing another picture, and so on *ad infinitum*.

So the main point being argued here is that the inner screen theory shown in **1.1** totally fails to explain how we can recognize the various objects and properties of objects in the visual scene. And the ability to achieve such recognition is anything but an immediate consequence of having a photographic representation. A television set has pixel images but it is precisely because it cannot decide what is in the scene from whence the images came that we would not call it a "visual perceiver." Devising a seeing machine that can receive a light image of a scene and use it to describe what is in the scene is much more complicated, a problem which is as yet unsolved for complex natural scenes.

The conclusion is inescapable: whatever the correct theory of seeing turns out to be, it must include processes quite different from the simple mirroring of the input image by simple point-by-point brain pictures. Mere physical resemblance to an input image is not an adequate basis for the brain's powers of symbolic visual scene description. This point is sometimes emphasized by saying that whereas the task of the eyes is forming images of a scene, the task of vision is the opposite: ***image inversion***. This means getting a description of the scene from images.

Images Are Not Static

For simplicity, our discussion so far has assumed that the eye is stationary and that it is viewing a stationary scene. This has been a convenient simplification but in fact nothing could be further from the truth for normal viewing. Our eyes are constantly shifted around as we move them within their sockets, and as we move our heads and bodies. And very often things in the scene are moving. Hence vision really begins with a stream of time-varying images.

Indeed, it is interesting to ask what happens if the eyes are presented with an unchanging image. This has been studied by projecting an image from a small mirror mounted on a tightly fitting contact lens so that whatever eye movement is made, the image remains stationary on the retina. When this is done, normal vision fades away: the scene disappears into something rather like a fog. Most visual processes just seem to stop working if they are not fed with moving images.

In fact, some visual scientists have claimed that vision is really the study of motion perception; all else is secondary. This is a useful slogan (even if an exaggeration) to bear in mind, particularly as we will generally consider, as a simplifying strategy for our debate, only single-shot stationary images.

Why do we perceive a stable visual world despite our eyes being constantly shifted around?

As anyone who has used a hand-held video camera will know, the visual scenes thus recorded appear anything but stable if the camera is jittered around. Why does not the same sort of thing happen to our perceptions of the visual world as we move our eyes, heads and bodies? This is an intriguing and much studied question. One short answer is that the movement signals implicit in the streams of retinal images are indeed encoded but then interpreted in the light of information about self-movements, so that retinal image changes due to the latter are cancelled out.

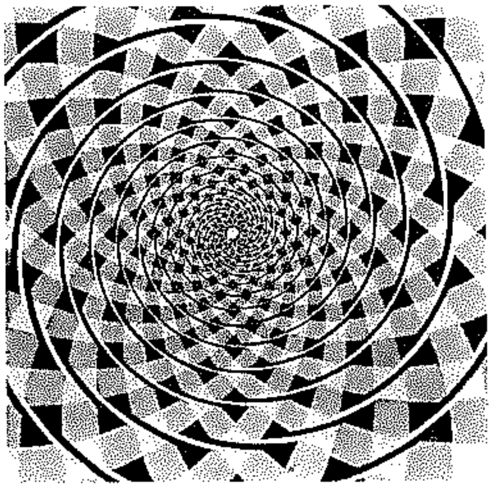
Visual Illusions and Seeing

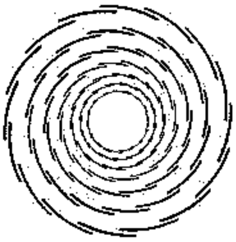
The idea that visual experience is somehow akin to photography is so widespread and so deeply rooted that many readers will probably not be convinced by the above arguments against the "inner screen" theory. They know that the eye does indeed operate as a kind of camera, in that it focuses an image of the world upon its light-sensitive retina.

An empirical way of breaking down confidence in continuing with this analogy past the eye and into the visual processes of the brain is to consider visual illusions. These phenomena of seeing draw attention to the fact that what we see often differs dramatically from what is actually before our eyes. In short, the non-photographic quality of visual experience is borne out by the large number and variety of visual illusions.

Many illusions are illustrated in this book because they can offer valuable clues about the existence of perceptual mechanisms devoted to building up an explicit scene description. These mechanisms operate well enough in most circumstances, but occasionally they are misled by an unusual stimulus, or one which falls outside their "design specification," and a visual illusion results. Richard Gregory is a major current day champion of this view (Gregory, 2009).

Look, for example, at **1.10**, which shows an illusion called ***Fraser's spiral***. The amazing truth is that there is no spiral there at all! Convince yourself of this by tracing the path of the apparent spiral with your finger. You will find that you return to your starting point. At least, you will if you are careful: the illusion is so powerful that it can even induce incorrect finger-tracing. But with due care, tracing shows that the picture is really made up of concentric circles. The spiral exists only in your head. Somehow the picture fools the visual system, which mistakenly provides a scene description incorporating a spiral even though no spiral is present. A process which takes concentric circles as input and produces a spiral as output can hardly be thought of as "photographic."



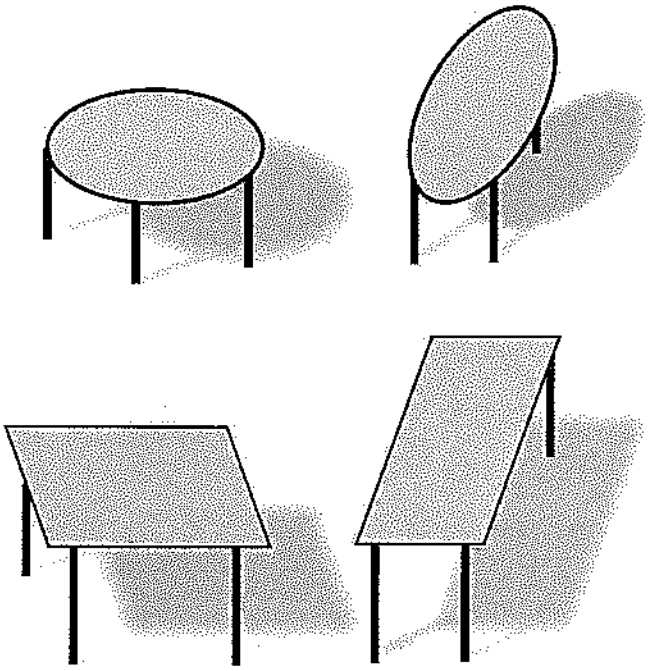


1.10 Fraser's spiral (above) This illusion was first described by the British psychologist James Fraser in 1908. Try tracing the spiral with your finger and you will find that there is no spiral! Rather, there are concentric circles composed of segments angled toward the center (left).

Another dramatic illusion is shown in **1.11**, which shows a pair of rectangles and a pair of ellipses. The two members of each pair have seemingly different shapes and sizes. But if you measure them with a ruler or trace them out, you will find they are the same. Incredible but true.

You might be wondering at this point: are such dramatic illusions representative of our everyday perceptions, or are they just unusual trick figures dreamt up by psychologists or artists? These illusions may surprise and delight us but are they really helpful in telling us what normally goes on inside our heads when we see the world? Some distinguished researchers of vision, for example, James Gibson, whose ***ecological optics*** approach to vision is described Ch 2, have argued that illusions are very misleading indeed.

But probably a majority of visual scientists would nowadays answer this question with a defi-

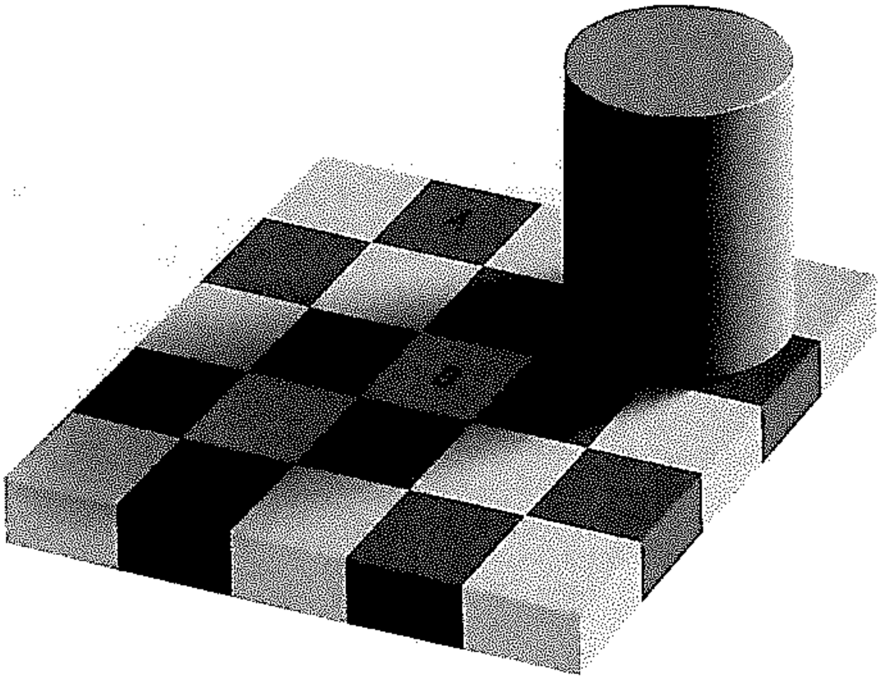


1.11 Size illusion
The rectangular table tops appear to have different dimensions, as do the elliptical ones. If you do not believe this then try measuring them with a ruler. Based on figure A2 in Shepard (1990).

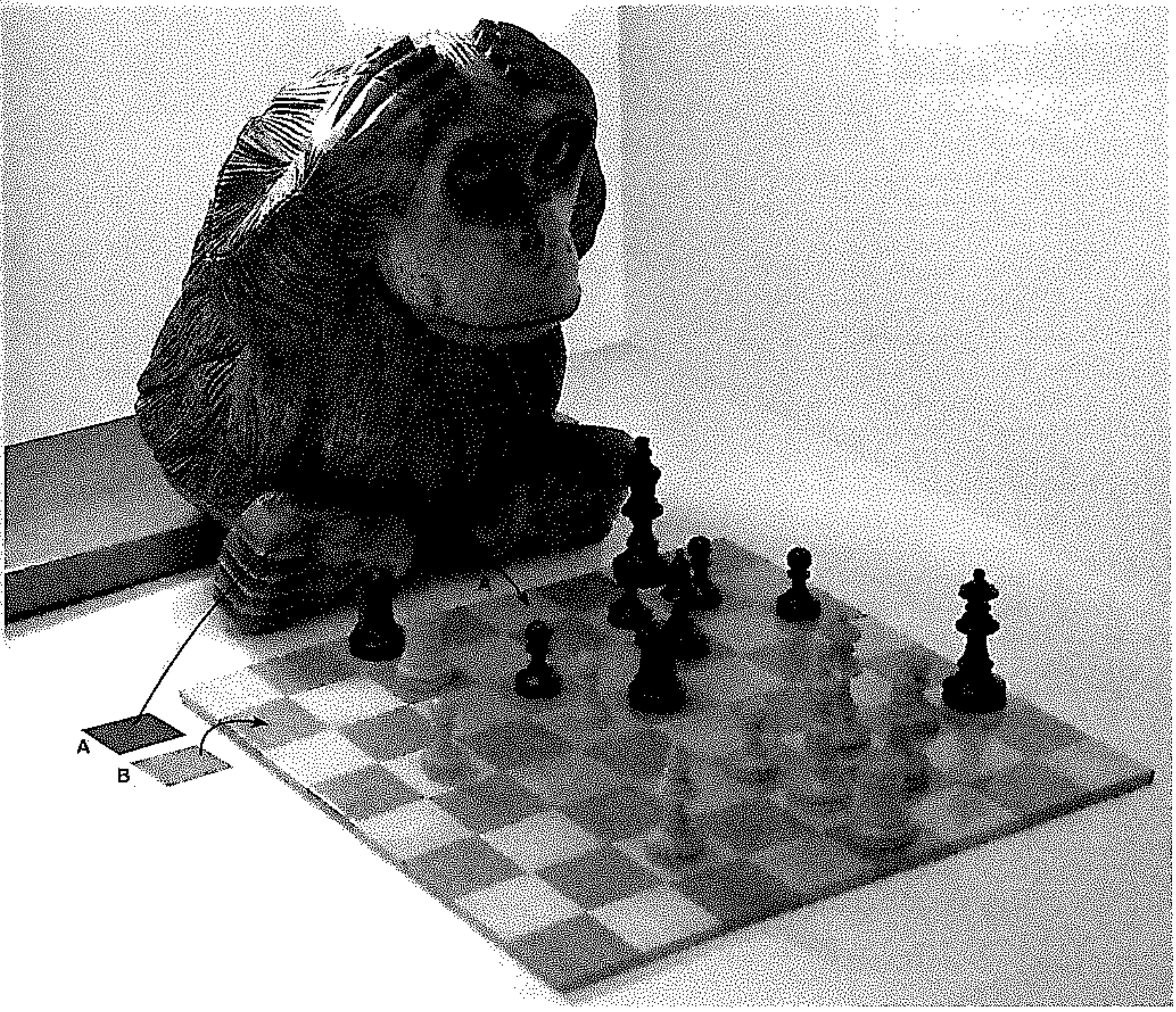
nite "yes." Visual illusions *can* provide important clues in trying to understand perceptual processes, both when these processes produce reasonably accurate perceptions, and when they are fooled into generating illusions. We will see how this strategy works out as we proceed through this book.

At this point we need to be a bit clearer about what we mean by a "visual illusion." We are using illusions here to undermine any remaining confidence you might have in the "inner screen" photographic-style theory of seeing. That is, illusions show that our perceptions often depart radically from predictions gained from applying rulers or other measuring devices to photographs.

But often visual illusions make eminently good sense if we regard the visual system as using retinal images to create representations of what really is "out there." In this sense, the perceptions are not illusions at all—they are faithful to "scene reality." A case in point is shown in **1.12**, in which a checkerboard of light and dark squares is cast in shadow. Unbelievably, the two squares picked out have the same intensity on-the-page in this computer graphic but they appear hugely different in lightness. This is best regarded not as an "illusion,"



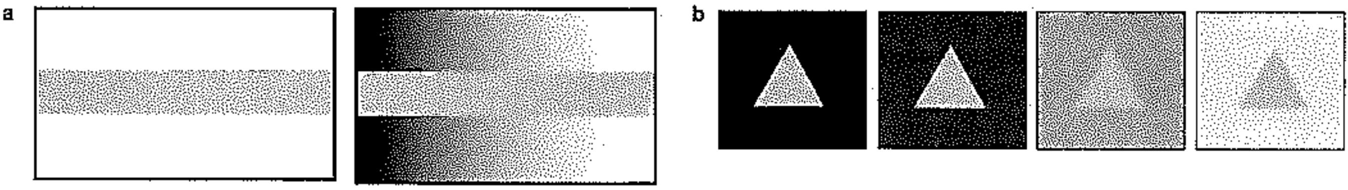
1.12 Adelson's figure
The squares labeled A and B have roughly the same gray printed on the page but they are perceived very differently. You can check their ink-on-the-page similarity by viewing them through small holes cut in a piece of paper. Is this a brightness illusion or is it the visual system delivering a faithful account of the scene as it is in reality? The different perceived brightnesses of the A and B squares could be due to the visual system allowing for the fact that one of them is seen in shadow. If so, the perceived outcomes are best thought of as being "truthful," not "illusory." See text. Courtesy E. H. Adelson.



1.13 A real-life version of the kind of situation depicted in the computer graphic in 1.12
Remarkably, the square in shadow labeled A has a lower intensity than the square labeled B, as shown by the copies at the side of the board. The visual system makes allowance for the shadow and to see what is "really there." [Black has due cause to appear distressed. After: 36 Bf5 Rxf5 37 Rxc8+ Kh7 38 Rh1, Black resigned. Following the forced exchange of queens that comes next, White wins easily with his passed pawn. Topalov vs. Adams, San Luis 2005.] Photograph by Len Hetherington.

strange though it might seem at first sight. Rather, it is an example of the visual system making due allowance for the shading to report on the true state of affairs (veridical).

The outcome in **1.12** is not some quirk of computer graphics, as illustrated in **1.13** in which the same thing happens from a photograph of a shaded scene.



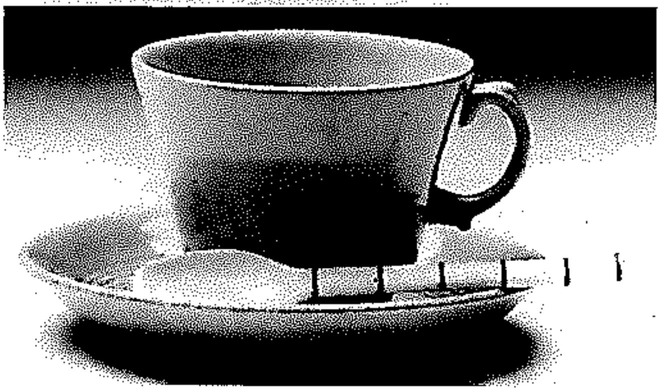
1.14 Brightness contrast illusions
a The two gray stripes have the same intensity along their lengths. However, the right hand stripe appears brighter at the end which is bordered by a dark ground, darker when adjacent to a light ground.
b The small inset gray triangles all have the same physical intensity, but their apparent brightnesses vary according to the darkness/lightness of their backgrounds. We discuss brightness illusions in Ch 16.

Other brightness "illusions" are shown in **1.14**. These also illustrate the slippery nature of what is to be understood by a "visual illusion." Figure **1.14a** could well be a case of making allowance for shading but **1.14b** doesn't fit that kind of interpretation because we do not see these figures as lying in shade. However, **1.14b** might be a case of the visual system applying, unconsciously, a strategy that copes with shading in natural scenes but when applied to certain sorts of pictures produces an outcome that surprises us because we don't see a shaded scene. This is not a case of special pleading because most visual processes are unconscious, so why not this one? On this argument, the varied perceptions of the identical-in-the-image gray triangles is "illusory" only if we are expecting the visual system to report the intensities in the image.

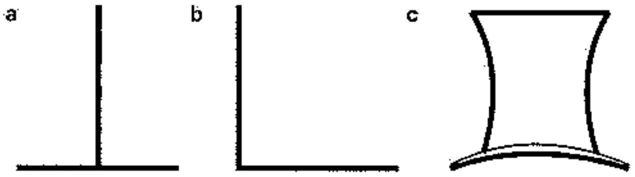
But, when you think about it, that doesn't make sense. The visual system isn't interested in reporting on the nature of the retinal image. Its task is to use retinal images to deliver a representation of what is *out there in the world*. The idea that vision is about seeing what is in retinal images of the world rather than in the world itself is at the root of the delusion that seeing is somehow akin to photography.

That said, if illusions are defined as the visual system getting it seriously wrong when judged against physically measured scene realities, then human vision is certainly prone to some illusions in this sense. These can arise for ordinary scenes, but go unnoticed by the casual observer. The teacup illusion shown in **1.15a** is an example. The photograph is of a perfectly normal teacup, together with a normal saucer and spoon. Try judging which mark on the spoon would be level with the rim of the teacup if the spoon was stood upright in the cup.

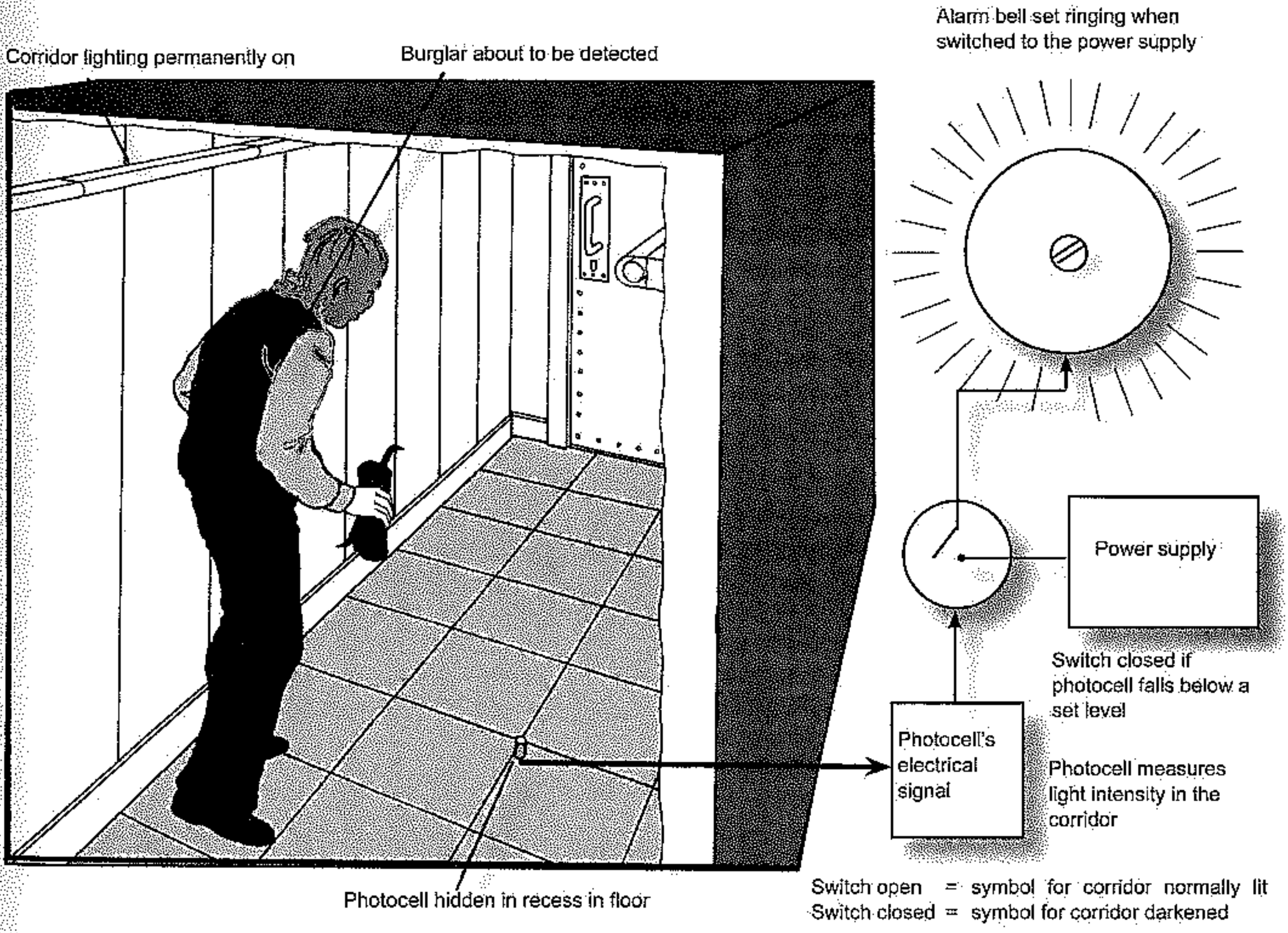
Now turn the page and look at **1.15b** (p.18). The illusory difference in the apparent lengths of the two spoons, one lying horizontally in the saucer and one standing vertically in the cup, is remarkable. Convince yourself that this percep-



1.15a Teacup illusion
Imagine the spoon stood upright in the cup. Which mark on the spoon handle would then be level with the cup's rim? Check your decision by inspecting **1.15b** on p.18.



1.16 The vertical-horizontal illusion
The vertical and horizontal extents are the same (check with a ruler). This effect occurs in drawings of objects, as in **c**, where the vertical and horizontal curves are the same length.


Alarm bell set ringing when switched to the power supply
Corridor lighting permanently on
Burglar about to be detected
Power supply
Switch closed if photocell falls below a set level
Photocell's electrical signal
Photocell measures light intensity in the corridor
Photocell hidden in recess in floor
Switch open = symbol for corridor normally lit
Switch closed = symbol for corridor darkened

1.17 A simple burglar alarm system operated by a photocell

tual effect is not a trick dependent on some subtle photography by investigating it in a real-life setting with a real teacup and spoon. It works just as well there as in the photograph. Real-world illusions like these are more commonplace than is often realized. Artists and craftsmen know this fact well, and learn in their apprenticeships, often the hard way by trial and error, that the eye is by no means always to be trusted. Seeing is *not* always believing—or shouldn't be.

This teacup illusion nicely illustrates the usual general definition of visual illusions—as perceptions which depart from measurements obtained with such devices as rulers, protractors, and light meters (the latter are called *veridical* measurements). Specifically, this illusion demonstrates that we tend to over-estimate vertical extents in comparison with horizontal ones, particularly if the vertical element bisects the horizontal one.

The simplest version of this effect, illustrated in **1.16a**, is known as the vertical-horizontal illusion. The effect is weaker if the vertical line does not bisect the horizontal as in **1.16b** but it is still present. It is easy to draw many realistic pictures containing the basic effect. The brim in **1.16c** is as wide as the hat is tall, but it does not appear that way. The perceptual mechanisms responsible for the vertical-horizontal illusion are not understood, though various theories have been proposed since its first published report in 1851 by A. Fick.

The illusions just considered are instances of ***spatial distortions***: vertical extents can be stretched, horizontal ones shortened, and so on. They are eloquent testimony to the fact that perceptions cannot be thought of as "photographic copies" of the world, even when it comes to a visual experience as apparently simple as that of seeing the length of a line.

Scene Descriptions Must Be Explicit

Explanations of various illusions will be offered in due course as this book proceeds. For the present, we will return to the theme of *seeing is representation*, and articulate in a little more detail what this means.

The essential property of a scene representation is that it makes some property of the scene *explicit* in a code of symbols. In the "inner screen" theory of **1.1**, the various brain cells make explicit the various shades of gray at all points in the image. That is, they signal the intensity of these grays in a way that is sufficiently clear for subsequent processes to be able to use them for some purpose or other, without first having to engage in more analysis. (When we say in this book a representation makes something explicit we mean: immediately available for use by subsequent processes, no further processing is necessary.)

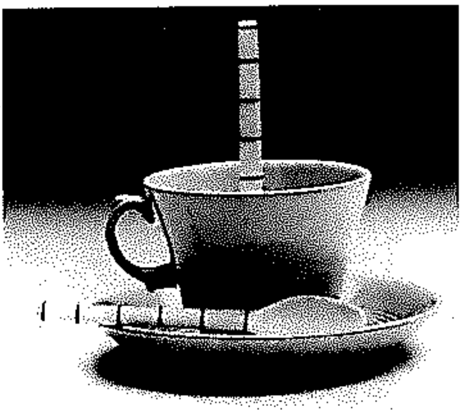
A scene representation then, is the result of processing an image of the scene in order to make attributes of the scene explicit. The simplest example we can think of that illustrates this kind of system in action is shown in **1.17**, perhaps the most primitive artificial "seeing system" conceivable—a burglar alarm operated by a photocell. The corridor is permanently illuminated, and when the intruder's shadow falls over the photocell detector hidden in the floor, an alarm bell is set ringing. Viewed in our terms, what the photocell-triggered alarm system is doing is:

1. Collecting light from a part of the corridor using a lens.

2. Measuring the intensity of the light collected—the job of the photocell;

3. Using the intensity measurement to build an explicit representation of the illumination in the corridor—*switch open* symbolizes *corridor normally lit* and *switch closed* symbolizes *corridor darkened*.

4. Using the symbolic scene description coded by the state of the switch as a basis for action—either ringing the alarm bell or leaving it quiet.



1.15b Teacup illusion (cont.)
The vertical spoon seems much longer than the horizontal one. Both are the same size with the same markings.

Step 3 requires some ***threshold*** level of photocell activity to be set as an ***operational*** definition of "corridor darkened." Technically, setting a threshold of this sort is called a ***non-linear*** process, as it transforms the linear output of the photocell (more light, bigger output) into a YES/NO ***category decision***.

Step 4 depends on the assumption that a darkened corridor implies "intruder." It would suffer from an "intruder illusion" if this assumption was misplaced, as might happen if a power cut stopped the light working.

The switch in the burglar alarm system serves as a symbol for "burglar present/absent" only in the context of the entire system in which it is embedded. This simple switch could be used in a different circuit for a quite different function. The same thing seems to be true of nerve cells. Most seem to share fundamentally similar properties in the way they become active, **1.3**, but they convey very different messages (code for different things, represent different things) according to the circuits of which they are a part. This type of coding is thus called ***place coding***, or sometimes ***value coding***, and we will see in later chapters how the visual brain uses it.

A primitive seeing system with similar attributes to this burglar detector is present in mosquito larvae: try creating a shadow by passing your hand over them while they are at the surface of a pond and you will find they submerge rapidly, presumably for safety using the shadow as warning of a predator.

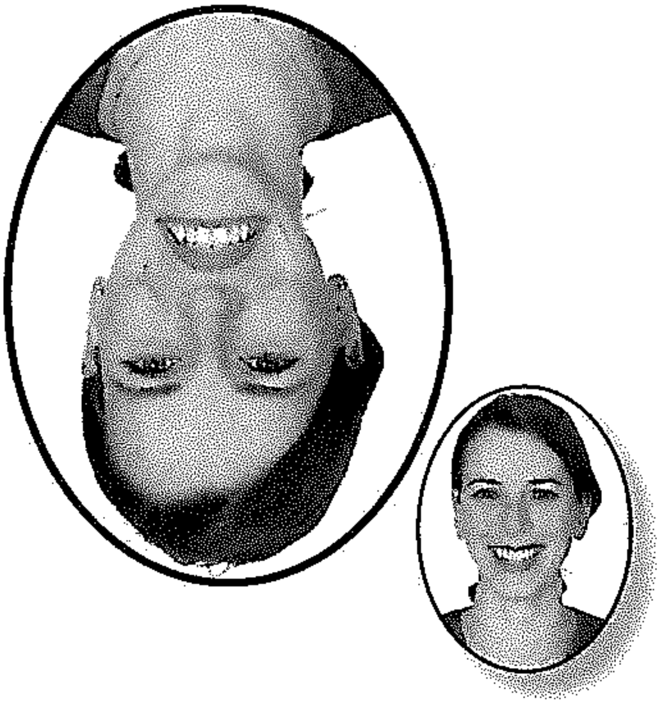
More Visual Tricks

The effortless fluency with which our visual system delivers its explicit scene representation is so beguiling that the skeptical reader might still doubt that building visual representations is what seeing is all about. It can be helpful to overcome this skepticism by showing various trick figures that catch the visual system out in some way, and reveal something of the scene representation process at work.

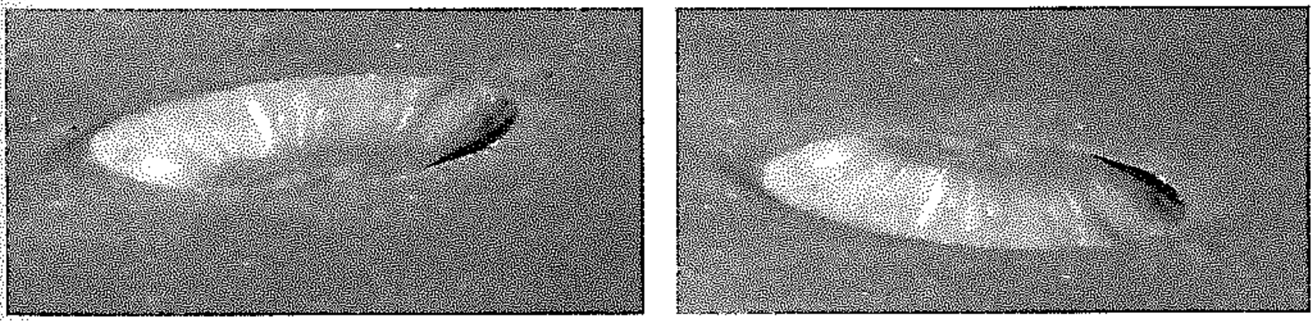
Consider, for example, the picture shown in **1.18.** It seems like a perfectly normal case of an inverted photograph of a head. Now turn it upside-down. Its visual appearance changes dramatically—it is still a head but what a different one.

These sorts of upside-down pictures demonstrate the visual system at work building up scene descriptions which best fit the available evidence. Inversion subtly changes the nature of the evidence in the retinal image about what is present in the scene, and the visual system reports accordingly. Notice too that the two alternative "seeings" of the photograph actually *look* different. It is not that we attach different verbal labels to the picture upon inversion. Rather, we actually *see* different attributes of the eyes and mouth in the two cases. The pattern of ink on the page stays the same, apart from the inversion, but the experience it induces is made radically different simply by turning the picture upside-down.

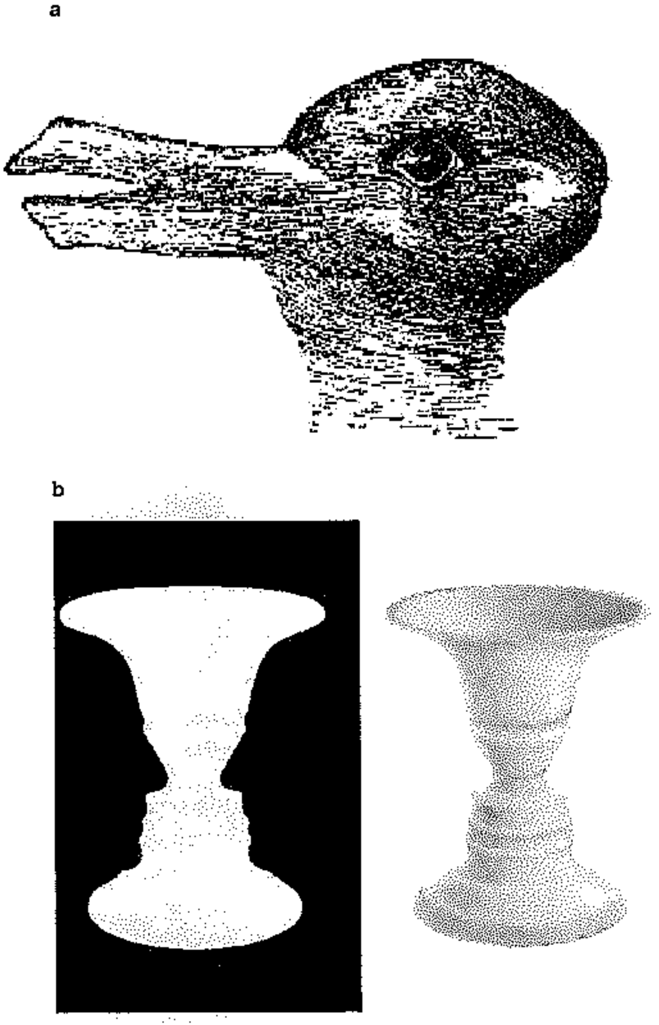
The "inner screen" theory has a hard time trying to account for the different perceptions produced by inverting **1.18.** The "inner screen" theorist wishes to reserve for his screen the job of representing the contents of visual experience. Fundamentally different experiences emerge upon inversion; therefore, fundamentally different contents must be recorded on the screen in each case. But it is not at all clear how this could be done. The "inner screen" way of thinking would predict that inversion should simply have produced a perception of the same picture, but upside-down. This is not what happens in **1.18** although it is what happens for pictures that lack some form of carefully constructed changes.



1.18 Peter Thompson's inverted face phenomenon
Turn the book upside-down but be ready for a shock.



1.19 Interpreting shadows
The picture on the right is is an inverted copy of the one on the left. Try inverting the book and you will see that the crater becomes a hill and the hill becomes a crater. The brain assumes that light comes from above, then it interprets the shadows to build up radically different scene descriptions (perceptions) of the two images. Courtesy NASA.



1.20 Ambiguous figures
a Duck or rabbit? This figure has a long history but it seems it was first introduced into the psychological literature by Jastrow in 1899.
b Vase or faces? From www.wpclipart.com.

Another example of the way inversion of a picture can show the visual system producing radically different scene descriptions of the same image is given in **1.19**. This shows a scene with a crater alongside one with a gently rounded hill. It is difficult to believe that they are one and the same picture, but turning the book upside-down proves the point. Why does this happen? It illustrates that the visual system uses an assumption that light normally comes from above, and given this starting point, the ambiguous data in the image are interpreted accordingly—bumps become hollows and vice versa on inversion.

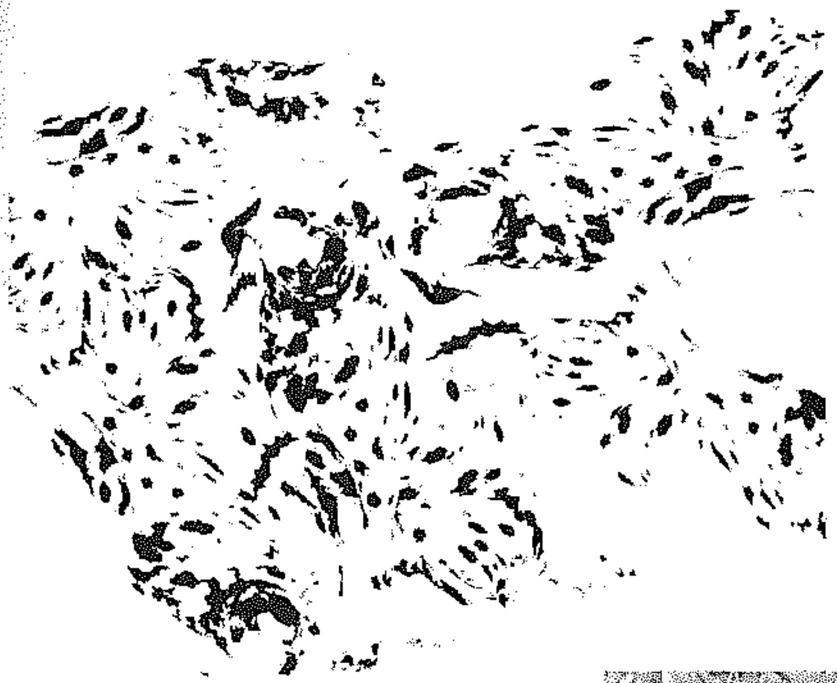
This is a fine example of how a visual effect can reveal a design feature of the visual system, that is, a principle it uses, an assumption it makes, in interpreting images to recover explicit scene descriptions. Such principles or assumptions are technically often called ***constraints.*** Identifying the constraints used by human vision is a critically important goal of visual science and we will have much to say about them in later chapters.

Figure/Ground Effects

Another trick for displaying the scene-description abilities of the visual system is to provide it with an ambiguous input that enables it to arrive at different descriptions alternately. **1.20** shows two classic ambiguous figures. The significance of ambiguous figures is that they demonstrate how different scene representations come into force at different times. The image remains constant, but the way we experience it changes radically. In **1.20a** picture parts on the left swap between being seen as ears or beak. In **1.20b** sometimes we see a vase as ***figure*** against its ***ground,*** and then at other times what was ground becomes articulated as a pair of faces—new figures.

Some aspects of the scene description do remain constant throughout—certain small features for instance—but the overall look of the picture changes as each possibility comes into being. The scene representation adopted thus determines the figure/ground relationships that we see. Just as with the upside-down face, it is not simply a case of different verbal labels being attached at different times. Indeed, the total scene description, including both features and the overall figure/ground interpretation, quite simply *is* the visual experience each time.

One last trick technique for demonstrating the talent of our visual apparatus for scene description is to slow down the process by making it more difficult. Consider **1.21** for example. What do you see there? At first, you will probably see little more than a mass of black blobs on a white ground. The perfectly familiar object it contains may come to light with persistent scrutiny but if you need help, turn to the end of this chapter to find out what the blobs portray.



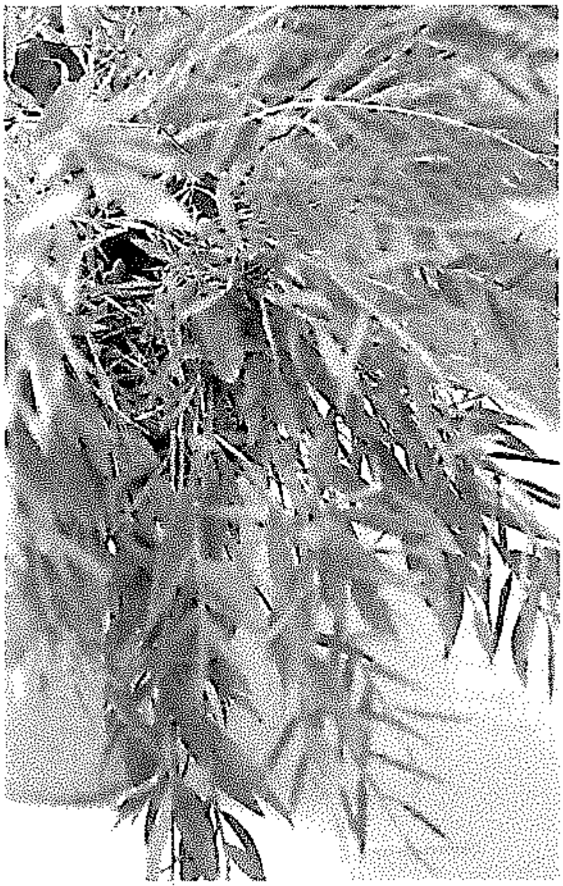
1.21 What do the blobs portray?
Courtesy Len Hetherington.

Once the hidden figure has been found (or, in our new terminology, we could say represented, described, or made explicit), the whole appearance of the pattern changes. In **1.21** the visual system's normally fluent performance has been slowed down, and this gives us an opportunity to observe the difference between the "photographic" representation postulated by the "inner screen" theory, and the scene description that occurs when we see things. The latter requires active interpretation of the available data. It is not "immediately given" and it is not a passive process.

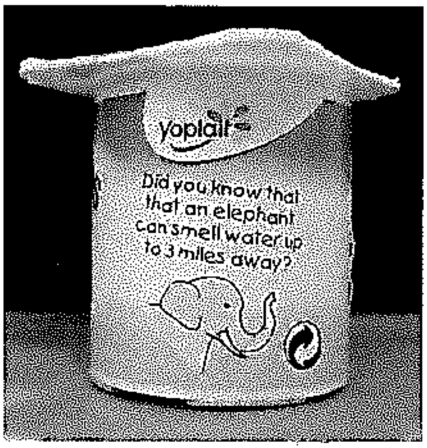
One interesting property of **1.21** is that once the correct scene description has been achieved, it is difficult to lose it, perhaps even impossible. One cannot easily return to the naive state, and experience the pictures as first seen.

Another example of a hidden-object figure is shown in **1.22.** This is not an artificially degraded image like **1.21** but an example of animal camouflage. Again, many readers will need the benefit of being told what is in the scene before they can find the hidden figure (see last page of this chapter for correct answers).

The use of prior knowledge about a specific object is called ***concept driven*** or ***top down processing.*** If such help is not available, or not used, then the style of visual processing is said to be ***data***

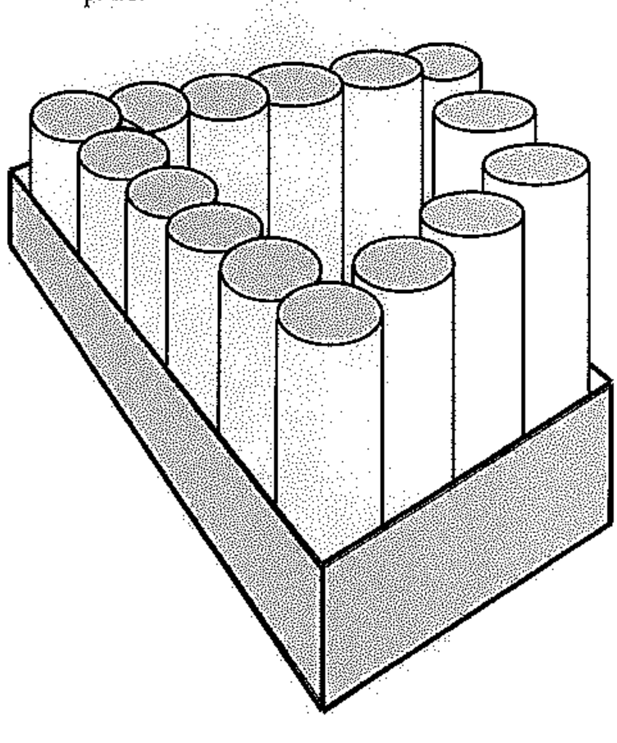


1.22 Animal camouflage
There are *two* creatures here. Can you find them?
Photograph by Len Hetherington.



1.23 Can you spot the error?
Thanks to S. Stone for pointing this out.

driven or ***bottom up***. An example of the way expectations embedded in concept driven processing can sometimes render us oblivious to what is "really out there" is shown by how hard it is to spot the unexpected error in **1.23**. For the answer, see p. 28.



1.24 Impossible pallisade
Imagine stepping around the columns, as though on a staircase. You would never get to the top (or the bottom). By J.P. Frisby, based on a drawing by L. Penrose and R. Penrose.

Three-Dimensional Scene Descriptions

So far we have confined our discussion of explicit scene descriptions to the problems of extracting information about objects from two-dimensional (2D) pictures. The visual system, however, is usually confronted with a scene in three dimensions (3D). It deals with this challenge magnificently and provides an explicit description of where the various objects in the scene, and their different parts, lie in space.

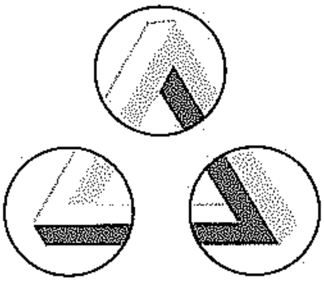
The "inner screen" theory cannot cope with the 3D character of visual perception: its representation is inherently flat. An attempt might be made to extend the theory in a logically consistent manner by proposing that the "inner screen" is really a 3D structure, a solid mass of brain cells, which represent the brightness of individual points in the scene at all distances. A kind of a brainware stage set, if you will.

It is doubtful whether complex 3D scenes could be re-created in brain tissue in a direct physical way. But even if this was physically feasible for 3D scenes, what happens when we see the "impossible pallisade" in **1.24**? (You may be familiar with the drawings of M. C. Escher, who is famous for having used impossible objects of this type as a basis for many technically intricate drawings.)

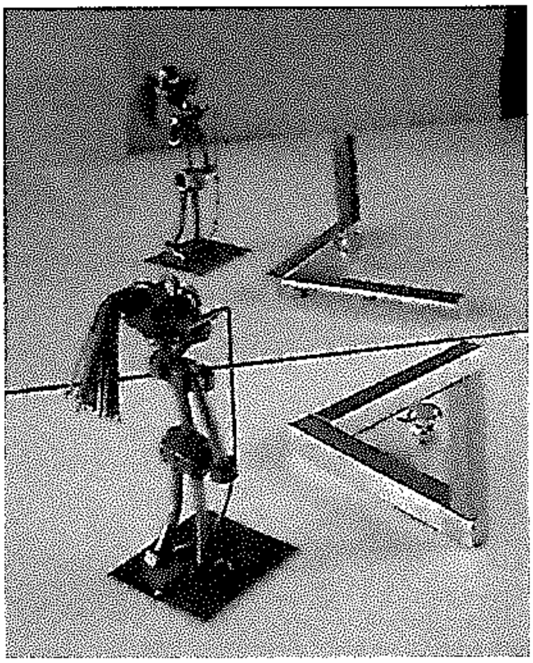
If this pallisade staircase is physically impossible, how then could we ever build in our brains a 3D physical replica of it? The conclusion is inescapable: we must look elsewhere for a possible basis for the brain's representation of depth (***depth*** is the term usually used by psychologists to refer to the distance from the observer to items in the scene being viewed, or to the different distances between objects or parts of objects).

What do impossible figures tell us about the brain's representation of depth? Essentially, they tell us that small details are used to build up an explicit depth description for *local* parts of the scene, and that finding a consistent representation of the entire scene is not treated as mandatory.

Just how the local parts of an *impossible triangle* make sense individually is shown in **1.25**, which gives an exploded view of the figure. The brain interprets the information about depth in each local part, but loses track of the overall description it is building up. Of course, it does not entirely lose



1.25 Impossible triangle
The triangle you see in the foreground in the photograph on the right is physically impossible. It appears to be a triangle only from the precise position from which the photograph was taken. The true structure of the photographed object is seen in the reflection in the mirror. The figure is included to help reveal the role of the mirror. This is a case in which the visual system prefers to make sense of local parts (the corners highlighted in the figure shown above), rather than making sense of the figure as a whole. Gregory (1971) invented an object of this sort. To enjoy diverse explorations of impossible objects, see Ernst (1996).



track of this global aspect; otherwise, we would never notice that impossible figures are indeed impossible. But the overall impossibility is a rather "cognitive" effect—a realization in thought rather than in experience that the figures do not "make sense."

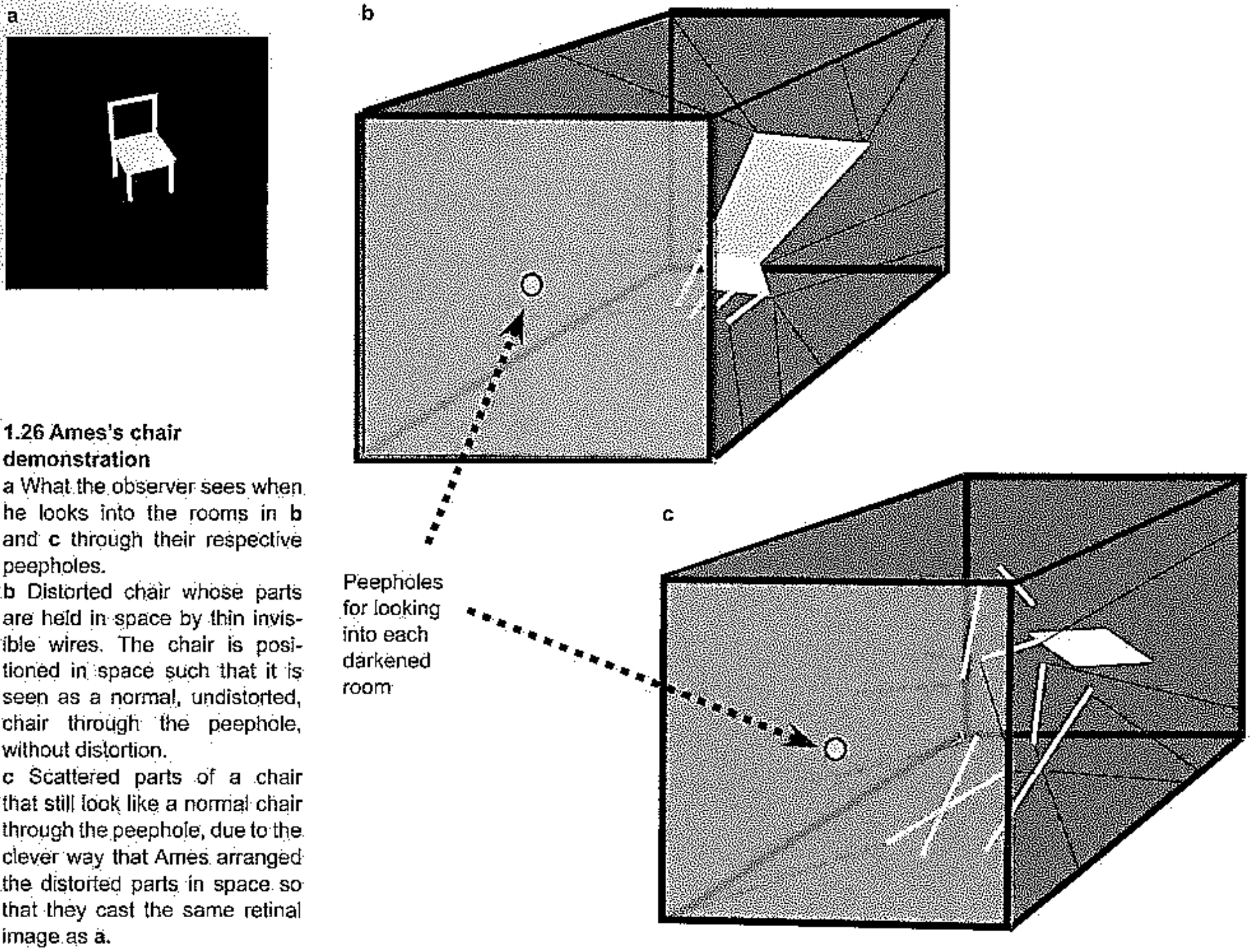
If the visual system insisted on the global aspect as "making sense" then it could in principle have dealt with the figures differently. For example, it could have "broken up" one corner of the impossible triangle and led us to see part of it as coming out toward us and part of it as receding. This is illustrated by the triangles in **1.25**.

But the visual system emphatically does not do this, not from a line drawing nor from a physical embodiment of the impossible triangle devised by Gregory. He made a 3D model of **1.25**, left. When viewed from just the right position, so that it presents the same retinal image as the line-drawing, then our visual apparatus still gets it wrong, and delivers a scene description which is impossible globally, albeit sensible locally. Viewing this "real" impossible triangle has to be one-eyed; otherwise, other clues to depth come into play and produce the physically correct global perception. (Two-eyed depth processing is discussed in detail in Ch 18.)

One interesting game that can be played with the trick model of the impossible triangle is to pass another object, such as one's arm, through the gap while an observer is viewing the model correctly aligned, and so seeing the impossible arrangement. As the arm passes through the gap, it seems to the observer that it slices through a solid object!

An important point illustrated by **1.25** is the inherent ambiguity of flat illustrations of 3D scenes. The real object drawn in **1.25**, left, has two limbs at very different depths: but viewing with one eye from the correct position can make this real object cast just the same image on the retina as one in which the two limbs meet in space at the same point.

This inherent ambiguity, difficult to comprehend fully because we are so accustomed to interpreting the 2D retinal image in just one way, is revealed clearly in a set famous demonstrations by Ames, shown in **1.26**. The observer peers with one eye through a peephole into a dark room and sees a chair, **1.26a**. However, when the observer is shown the room from above it becomes apparent that the real object in the room is *not* the chair seen through the peephole. In the example shown in **1.26b**, the object is a distorted chair suspended in space by invisible wires, and in **1.26c** the room contains an odd assemblage of luminous lines, also suspended in space by wires. The collection of parts is cunningly arranged in each case to produce



1.26 Ames's chair demonstration
a What the observer sees when he looks into the rooms in **b** and **c** through their respective peepholes.
b Distorted chair whose parts are held in space by thin invisible wires. The chair is positioned in space such that it is seen as a normal, undistorted, chair through the peephole, without distortion.
c Scattered parts of a chair that still look like a normal chair through the peephole, due to the clever way that Ames arranged the distorted parts in space so that they cast the same retinal image as **a**.

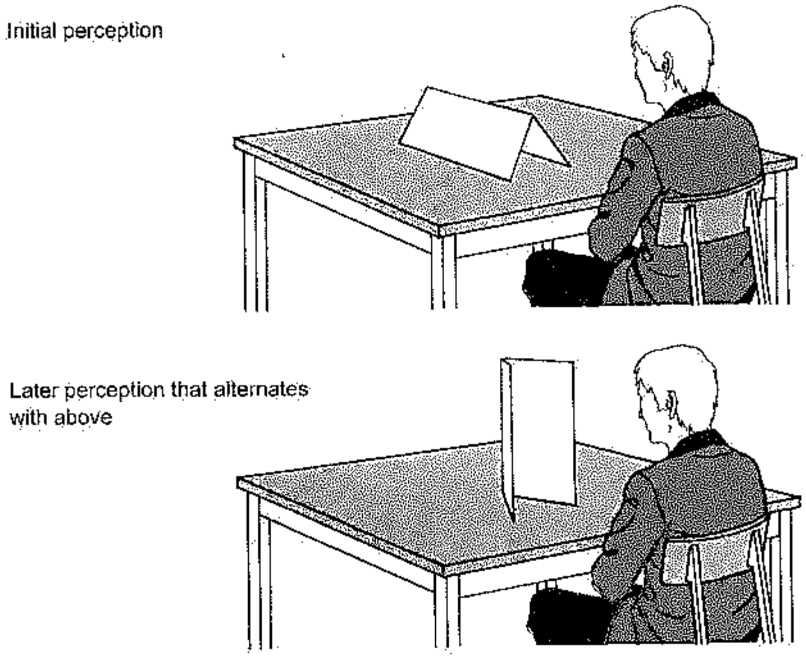
a retinal image which mimics that produced by the chair when viewed from the intended vantage point. In the most dramatic example, the lines are not formed into a single distorted object, but lie in space in quite different locations—and the "chair seat" is white patch painted on the wall.

The point is that the two rooms have things within them which result in a chair-like retinal image being cast in the eye. The fact that we see them as the same—as chairs—is because the visual system's design exploits the assumption that is "reasonable" to interpret retinal information in the way which normally yields perceptions that would be valid from diverse viewpoints. It is "blind" to other possibilities, but that should not deceive us—those possibilities do in fact exist.

Another way of putting this is to say that the Ames's chair demonstrations reveal that the visual system (along with a typical computer vision system) utilizes what is called the ***general viewpoint constraint.*** A normal chair appears as a chair from all viewpoints, whereas the special cases used by Ames can be seen as chairs from just one special vantage point. The general viewpoint constraint justifies visual processes that would yield stable structural interpretations as vantage point changes.

The general viewpoint constraint can be embedded in bottom up processing. It is not necessary to invoke top down processing in explaining the Ames chair demonstrations—that is, knowing the shape of normal chairs, and using this knowledge to guide the interpretation of the retinal image.

Normal scenes are usually interpreted in one way and one way only, despite the retinal image information ambiguity just referred to. But it is possible to catch the visual system arriving at



1.27 Mach's illusion
With one eye closed, try staring at a piece of folded paper resting on a table (upper). After a while it suddenly appears not as a tent but as a raised corner (lower). If the viewer moves while maintaining the illusory depth perception, then the object will appear to move with the viewer's movement.

different descriptions of an ambiguous 3D scene in the following way. Fold a piece of paper along its mid-line and lay it on a table, as in **1.27**. Stare at a point about mid-way along its length, using just one eye. Keep looking and you will suddenly find that the paper ceases to look like a tent as it "should" do, and instead looks like a corner viewed from the inside. The effect is remarkable and well worth trying to obtain.

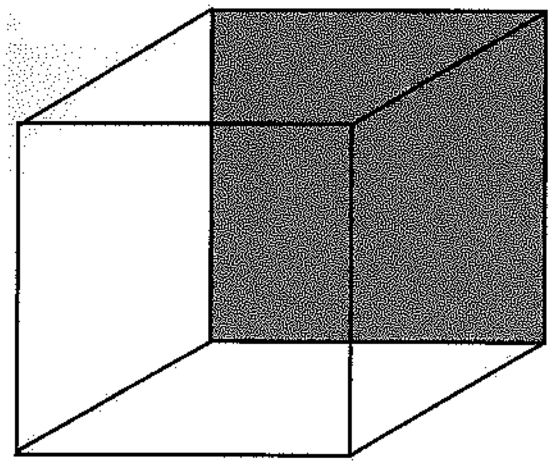
The point is that both "tent" and "corner" cast identical images on the retina, and the visual system sometimes chooses one interpretation, sometimes another. It could have chosen many more of course, and the fact that it confines itself to these two alternatives is itself interesting.

Another famous example of the same sort of alternation, but from a 2D drawing rather than from a 3D scene, is the Necker cube, **1.28**.

Conclusions

Perhaps enough has been said by now to convince even the most committed "inner screen" theorist that his photographic conception of seeing is wholly inadequate. Granted then that seeing is the business of arriving at explicit scene representations, the problem becomes: how can this be done?

It turns out that understanding how to extract explicit descriptions of scenes from retinal images is an extraordinarily baffling problem, which is one reason why we find it so interesting. The problem is at the forefront of much scientific and technological research at the present time, but it still remains largely intractable. Seeing has puzzled philosophers and scientists for centuries, and it continues to do so. To be sure, notable advances have been made in recent years on several fronts



1.28 Necker cube
Prolonged inspection results in alternating perceptions in which the shaded side is sometimes seen nearer, sometimes farther.

within psychology, neuroscience, and machine image-processing, and many samples of this progress will be reviewed in this book. But we are still a long way from being able to build a machine that can match the human ability to read handwriting, let alone one capable of analyzing and describing complex natural scenes.

This is so despite multi-million dollar investments in the problem because of the immense industrial potential for good processing systems. Think of all the handwritten forms, letters, etc. that still have to be read by humans even though their contents are routine and mundane, and all the equally mundane object handling operations in industry and retailing.

Whether we will witness a successful outcome to the quest to build a highly competent visual robot in the current century is debatable, as is the question of whether a solution would impress the ordinary person.

A curious fact that highlights both the difficulties inherent in understanding seeing and the way we take it so much for granted is that computers can already be made which are sufficiently "clever" to beat the human world champion at chess. But computers cannot yet be programmed to match the visual capacities even of quite primitive animals. Moves are fed into chess playing computers in non-visual ways. A computer vision system has not yet been made that can "see" the chessboard, from differing angles in variable lighting conditions for differing kinds of chess pieces—even though the computer can be made to play chess brilliantly.

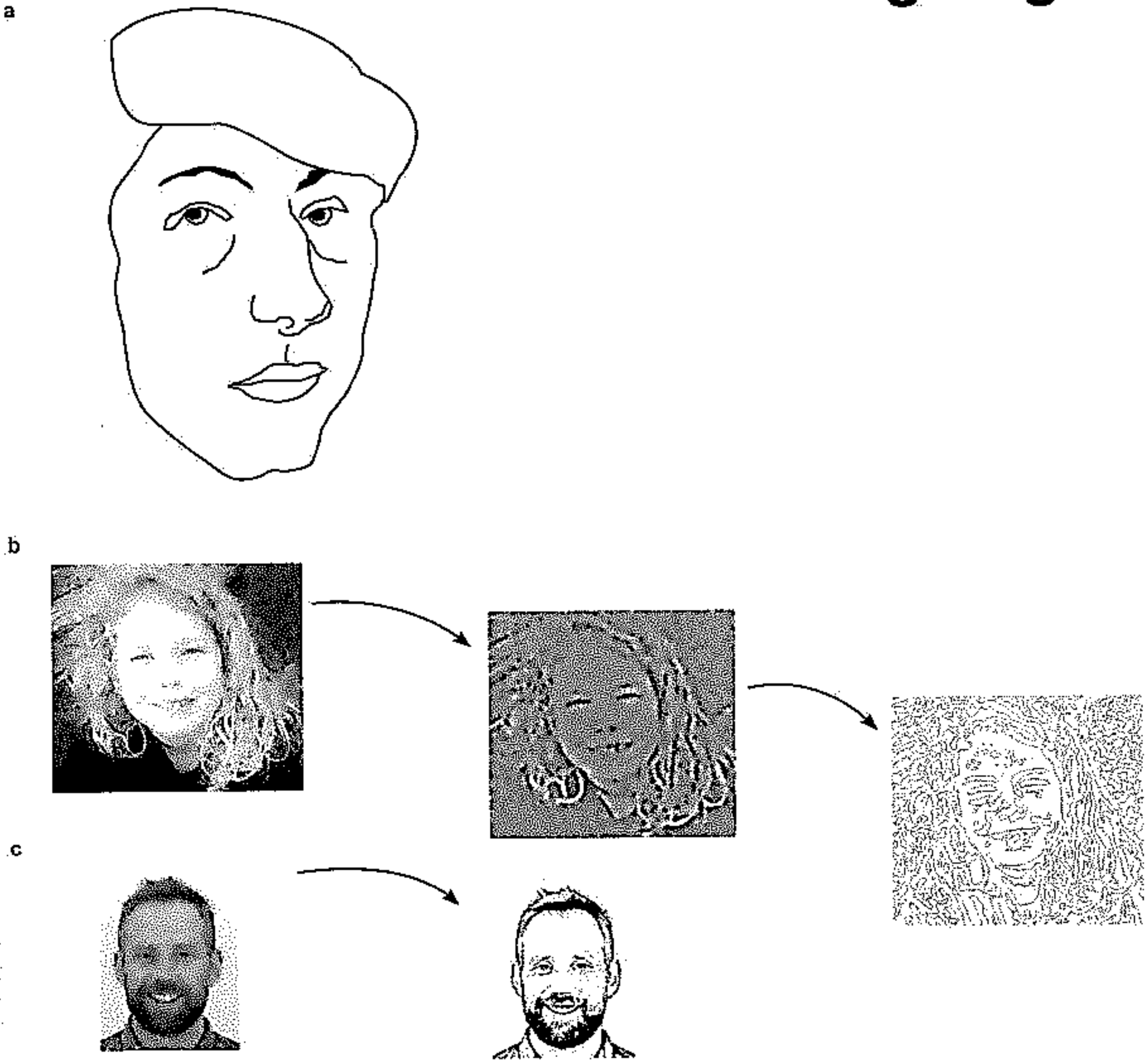
Even so, most people would probably be more impressed with a world-class chess-playing computer than they would be with a good image-processor, despite the fact that the former has been realized whereas the latter remains elusive. It is one of our prime objectives to bring home to you why the problem of seeing remains so baffling. Perhaps by the end of the book you will have a greater respect for your magnificent visual apparatus.

Meanwhile, we have said enough in this opening chapter to make abundantly clear that any attempt to explain seeing by building representations which simply mirror the outside world by some sort of physical equivalence akin to photography is bound to be insufficient. We do not see our retinal images. We use them, together with prior knowledge, to build the visual world that is our representation of what is "out there." We can now finally dispatch the "inner screen" theory to its grave and concentrate henceforth on theories which make ***explicit scene representations*** their objective.

In tackling this task, the underlying theme of this book will be the need to keep clearly distinct three different levels of analysis of seeing. Ch 2 explains what they are and subsequent chapters will illustrate their nature using numerous examples. We hope that by the time you have finished the book that we will have convinced you of the importance of distinguishing between these levels when studying seeing, and that you will have a good grasp of many fundamental attributes of human and, to a lesser extent animal, vision.

5

Seeing Edges



5.1 Edges in art and in computer vision
a Sketch of some of the main contours used in a self-portrait by Albrecht Dürer, illustrating how artists use lines to mark the boundaries of objects and their parts, such as here the eyes, nose, mouth, etc. The visual system seems to create similar edge-based representations, marking significant entities in retinal images.
b A processing sequence from image (left), via convolution using a circularly symmetric receptive field (center), to a representation of many of the edges in the image (right). This is the processing sequence described in this chapter.
c From image to sketch, by the computer vision system of Bruce Gooch, Erick Reinhard, and Amy Gooch (2004), reproduced with permission. The early stage of their system starts with a similar sequence to that shown in **b**, but then adds other processing stages.

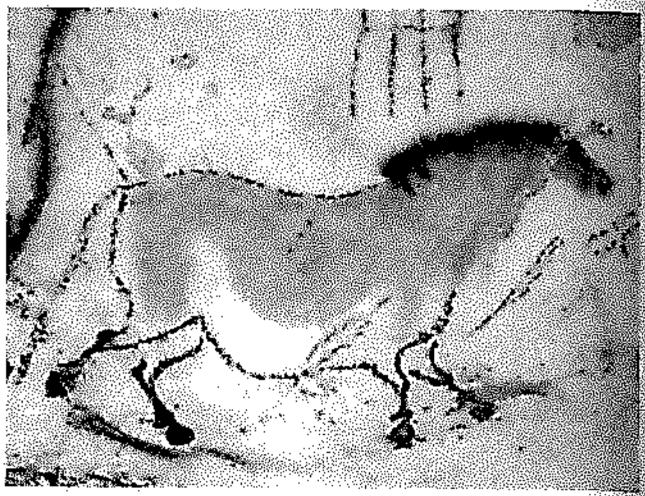
Take a look around the scene before you and you will have no trouble seeing ***edges*** formed by the boundaries of objects, of surfaces, of surface markings, and of shadows. This ability to see edges must reflect the fact that there are edge feature representations somewhere inside our heads.

Artists often use lines to pick out object edges, **5.1**, and they have known about the importance of edges in seeing since the beginning of art, as cave drawings testify, **5.2**.

It is of great theoretical interest to the science of seeing that line drawings of objects do not need to be complete for us to be able to recognize objects, **5.3** (see also Ch 8). In its search for object boundaries, human vision has ways to get by with partial evidence. The term ***illusory contour*** is used for an edge that is seen where none exists in the image (Ch 16). The importance of being able to cope with breaks in edges soon became apparent in the early days of computer vision. It was found, to the astonishment of investigators, that sometimes the pixel values in their grey level images simply did not have changes marking edges which they themselves could see when looking at the original images. This shocking realization led some of the pioneers of computer vision to postpone working from natural images and instead use hand-made edge maps for exploring various visual processes.

Why read this chapter?

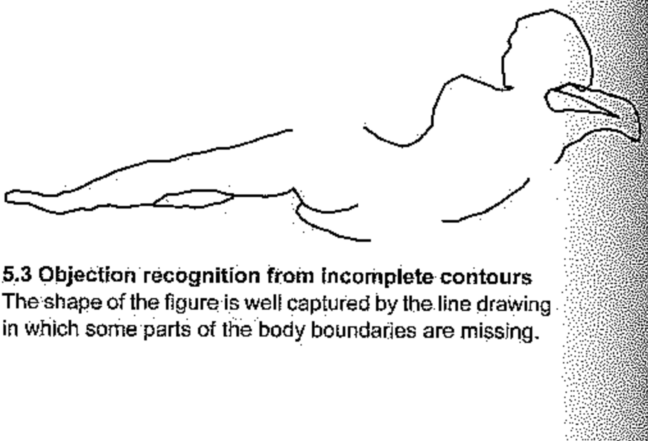
If we are to understand human vision it will be vital to understand the early stages of image processing in the visual pathway. We introduce this topic with Marr and Hildreth's computational theory of edge detection. A ubiquitous problem in finding edges is that images are imperfect; they contain glitches that are referred to as *noise*. The theory tackles the problem of noise by blurring images, using a process called *convolution*. In essence, this consists of applying a particular kind of *operator* (cf. receptive field in biological vision) all over the image. This operator has a gaussian profile because the computational theory specifies that this is the shape that optimally combines smoothing away noise with not disturbing too greatly where edges are to be found in the convolved image. The next step is to find regions in the image where there are abrupt changes in image intensities, because this is where the edges are located. This requires measuring intensity *gradients* and/or *changes in gradients*. These are called the *first* and *second derivatives* respectively. Various biologically plausible algorithms are described for implementing the theory. These algorithms involve operators that are remarkably similar to the receptive fields possessed by cells in the retina and the striate cortex. Using the task of edge detection, the chapter illustrates the value of distinguishing between the computational theory, algorithm, and hardware levels when trying to understand complex information processing systems.



5.2 Cave painting
From Lascaux, France. PD-Art.

Brain representations for edges cannot be all there is to seeing, of course, because we are able to describe scene characteristics much more complicated than edges. Even so, edge feature representations might be immediately useful for guiding a grasping action around an object. They can also serve as an important first step on the way to more complex perceptual tasks, such as object recognition or depth perception, as we will see.

How should edge detection be tackled? Bars are a species of edge feature, and we considered in Ch 3 the task of designing a bar detector. This task proved surprisingly tricky but we showed how bar orientation could be computed from the outputs



5.3 Objection recognition from incomplete contours
The shape of the figure is well captured by the line drawing in which some parts of the body boundaries are missing.

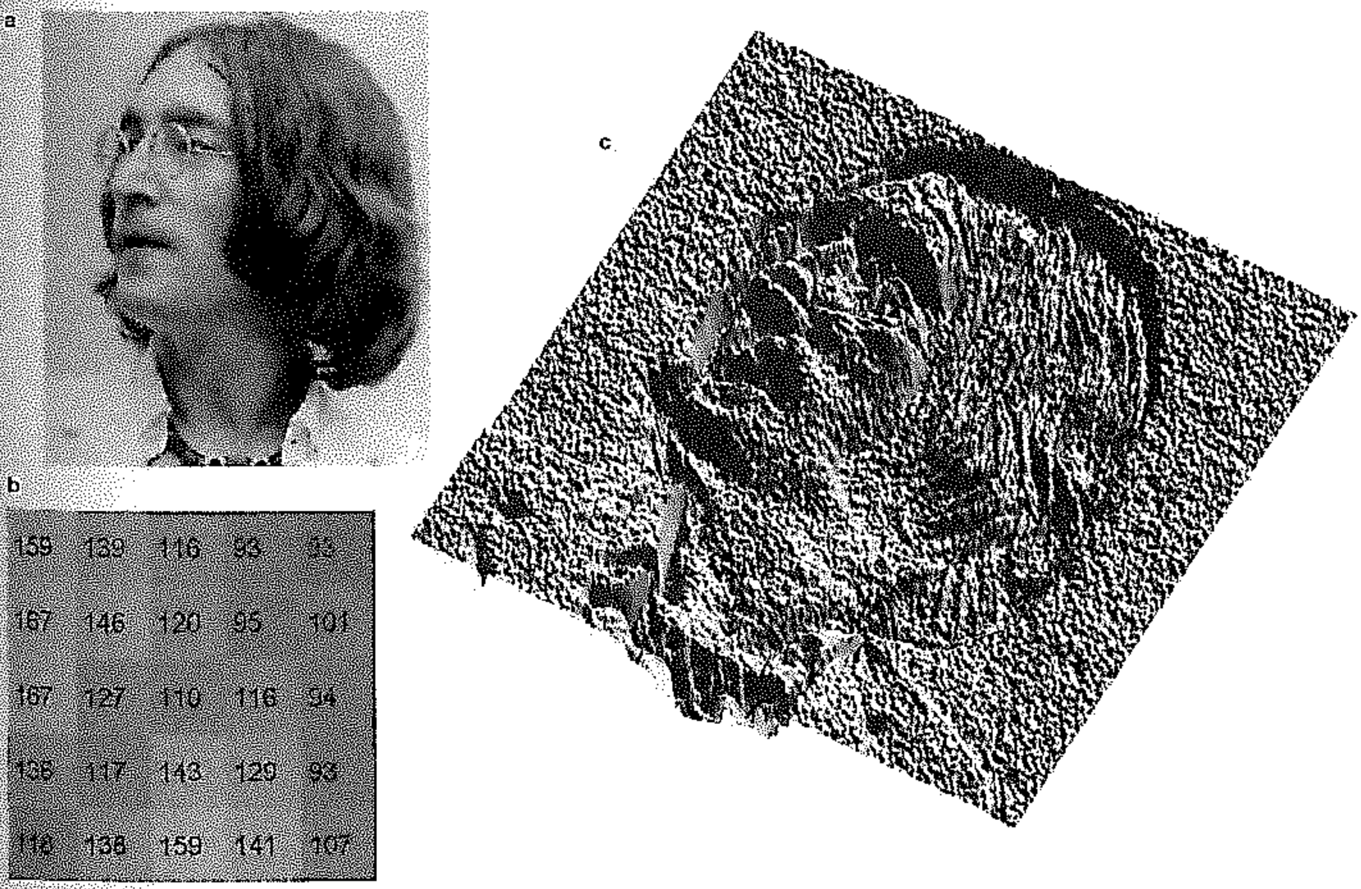
of a group of simple cells whose preferred stimuli were bars of different orientations. But there are many types of features other than bars and much more to the edge detection problem than just computing edge orientation. It is time to have a closer look at the seeing problem of edge detection.

Edges in Images

In studying seeing, it is always important to distinguish clearly between scenes and the images of those scenes. So the first question to ask in analyzing the problem of edge detection is: how do edges in scenes show up in images?

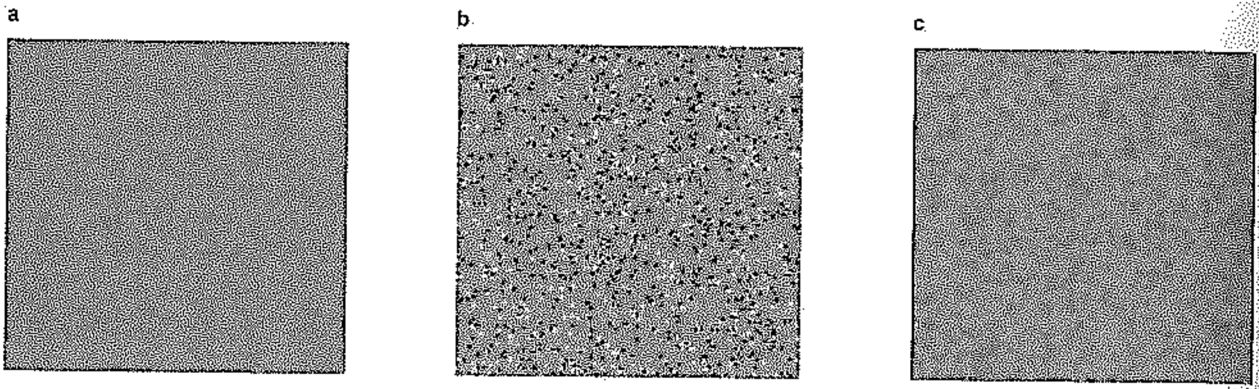
An image of John Lennon is shown in 5.4c. The pixel intensities of this image are shown as a hilly landscape in **5.4c**, in which height is used as a way of representing intensity. You may be surprised how broken up and cluttered this landscape appears. Can our clear percepts of the scene in front of the eyes really start from such a basis? The answer is *Yes*. Natural images usually are as messy as this landscape suggests.

Part of the messiness in **5.4b** is caused by the many edges from the hairs on Lennon's head. But look at the background in the original image and in the landscape. The former appears a pretty flat grey to our eyes whereas the latter is surprisingly "bubbly." Some of this "bubblyness" is caused by images having been captured with electronic light detectors that are a little "noisy." You can think of this ***noise*** as similar to the "snow" seen on a poorly tuned television. Receptors in the retina also suf-



159	139	116	93	93
167	146	120	95	101
167	127	110	116	94
136	117	143	129	93
118	138	159	141	107

5.4 Pixel Intensities as a landscape
a Photograph of John Lennon.
b Sample of pixels from a greatly magnified image. Each square represents one pixel whose intensity (gray level) is given as a number.
c Pixel intensities plotted as a 3D graph, which shows them as a hilly landscape in which the valleys indicate low intensities and the peaks high intensities.



5.5 Image noise cleaning
a Original test image. b Original image with added "salt and pepper" random noise. c The result of smoothing b by local 3x3 neighborhood averaging.

fer from random noise fluctuations, so noise is a problem for human as well as computer vision. Our first problem to be solved is thus clear. We need to find a principled way to eliminate image noise, or at least reduce it, prior to seeking edges in images. This ensures that any edges found can be relied upon to reflect scene edges and not spurious noise-created image intensity changes.

Computational Theory: Getting Rid of Image Noise

Analyzing this noise cleaning task properly would mean developing a deep understanding of where the noise comes from. This would be needed to develop a principled way of getting rid of it or sidestepping it in some way. In short, and as usual in studying seeing, a full ***task analysis*** would be needed to devise a good computational theory for getting rid of noise.

We will not explore noise sources in any detail, as that would lead us into some complex details about various visual neurons. Instead, we will consider just noise in receptors and we will proceed on the basis of a very simple idea. We will assume that the main sources of receptor noise are spatially random. What this means is that the noise in neighboring receptors is assumed to be independent. Think of this as each receptor being subject to its own "private" noise, so that noise fluctuations in one receptor cannot be predicted from those in its neighbors.

This assumption (another example of what is technically called a *constraint*) allows a neat trick for getting rid of noise: exploit the "law of averages" to cancel out a lot of the noise. That is, take the average of the activities in neighboring receptors and high noise in one will tend, due to the noise independence assumption, to be canceled out by low noise in another. Thus the total receptor noise will tend to average out to zero. [This scheme is not aimed at noise elsewhere in the visual system, particularly low spatial frequency noise.]

The price to be paid for using this constraint is that taking neighborhood averages blurs the grey level description a little, as we will see. But it does lead to smoother, less noisy, images and the penalty of some blur turns out to be worth paying. Indeed, we will find that using a range of images blurred to different extents is an important aspect of edge detection, because it is part of the process of finding edges of different types (from sharp through to fuzzy).

You might be thinking that the averaging scheme just proposed is a bit extreme. Wouldn't it be better to give most weight to the central receptor under consideration than to weight it equally with its neighbors? That is, instead of substituting the intensity recorded in each receptor with the average value of that receptor and its neighbors, how about using a scheme in which the neighbors somehow contribute relatively less weight to the end result? It turns out that this is exactly the right thing to do. But before explaining this important refinement, we will describe the equal-weighting idea first because it is simpler for conveying some core ideas.

An Algorithm for Image Noise Cleaning

How can neighborhood averaging be implemented? A possible computer vision ***algorithm*** *(*or *procedure)* is as follows:

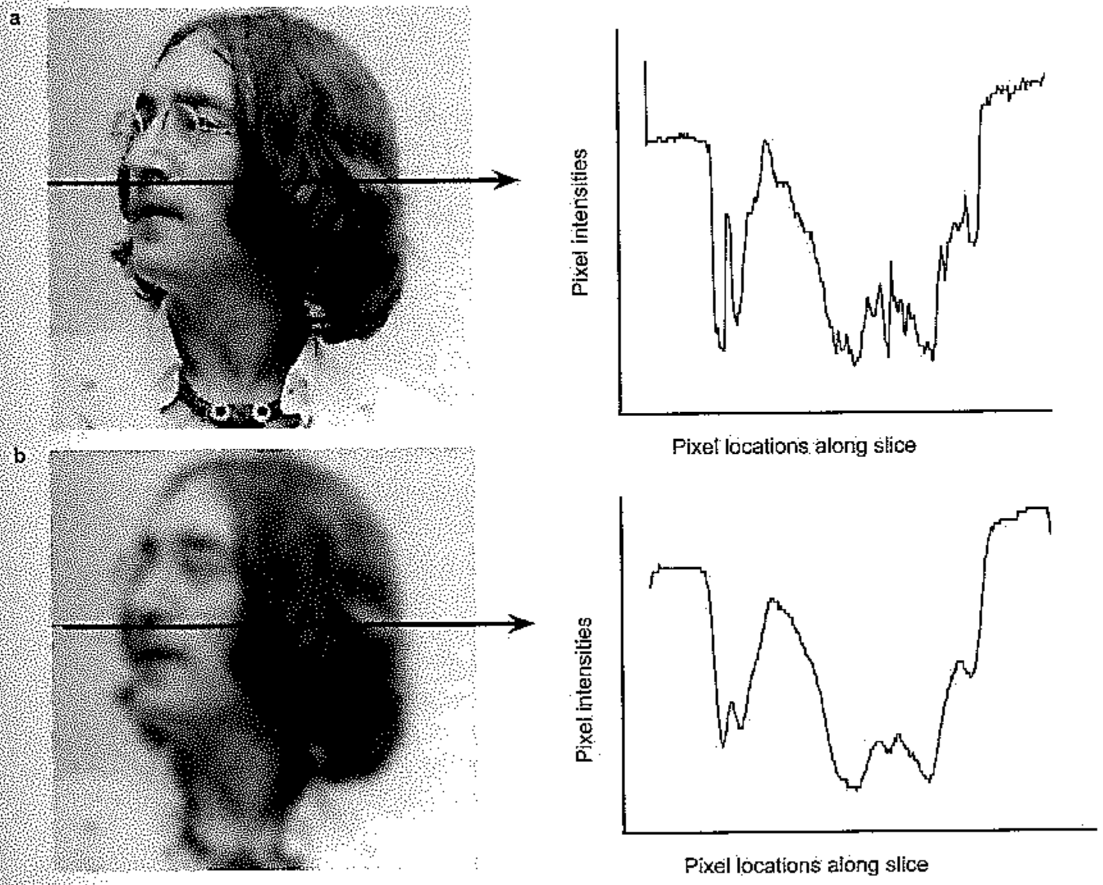
Step 1 For each pixel, add up the pixel intensities of its immediate neighbors, and then add this total to the intensity of the pixel in question. Call this the neighborhood total.

Step 2 For each pixel, divide its neighborhood total by the number of pixels contributing to that total, to calculate a mean (average) intensity. Substitute this mean for the original pixel intensity.

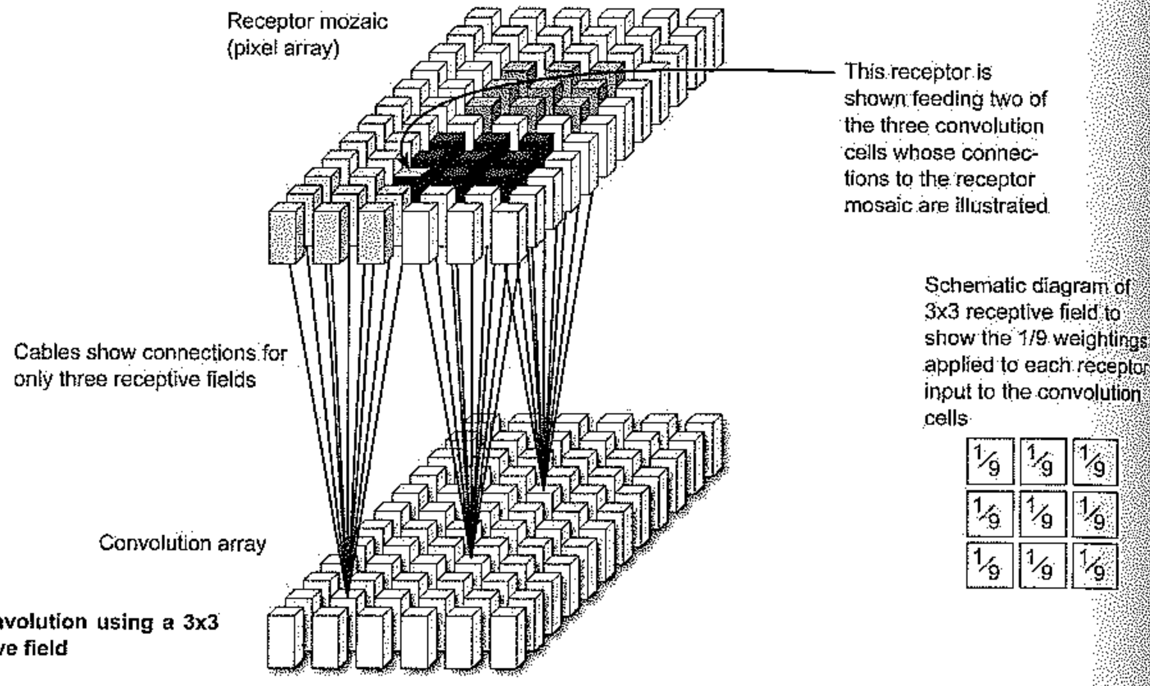
In short, add up all the pixel intensities in a region and divide by the number of pixels in that region. The region can be of various sizes but for the present we will use 3×3 patches of pixels (that is, each pixel plus its 8 closest neighbors).

Does this kind of neighborhood averaging work in getting rid of noise? The idea can be tested by taking a uniform grey image, **5.5a**, and adding noise randomly to each pixel, **5.5b**. To highlight the basic ideas, much more noise has been added than would be expected in either artificial or biological images. Running the 3×3 local averaging algorithm on this very noisy image shows that it reduces the noise quite a lot, **5.5c**, although it does not get rid of it all. Nevertheless, this test demonstrates that averaging can help reduce noise for independent receptor noise sources.

But what about local averaging on natural images? It produces a smoother image for the John Lennon picture, as can be seen in **5.6**.



5.6 Image smoothing
a Original Lennon image with alongside it the intensity profile of a one-dimensional slice indicated by the horizontal line.
b The same after the image has been smoothed using 3x3 local neighborhood averaging. Notice that both profiles show image intensity variations at different scales from steep to shallow.


Receptor mozaic (pixel array)
This receptor is shown feeding two of the three convolution cells whose connections to the receptor mosaic are illustrated
Schematic diagram of 3x3 receptive field to show the 1/9 weightings applied to each receptor input to the convolution cells
1/9 1/9 1/9
1/9 1/9 1/9
1/9 1/9 1/9
Cables show connections for only three receptive fields
Convolution array

5.7 Convolution using a 3x3 receptive field

Convolution

The procedure just outlined is all very well for computers, as they are good at arithmetical operations, but how might the brain do the averaging job with only neurons at its disposal? Can they be made to "perform" the required arithmetic using a method that fits how they work? The answer is yes, but to explain it entails introducing a generally useful procedure for computers as well as brains, called ***convolution***.

Convolution is illustrated in **5.7**. At the top is shown schematically a ***receptor mosaic*** (roughly equivalent to an array of pixels in computer vision, Ch 1) whose activities encode image grey levels. This receptor mosaic can be thought of as a highly simplified model of a small region of the receptor mosaic of a human eye. The receptors send their outputs into a second array of elements. We will call these ***convolution cells***, and they are arranged in a ***convolution array***. The latter array carry out a function similar to that performed by the ***bipolar cells*** in the retina, as we will see in Ch 6.

Each convolution cell in **5.7** receives inputs from 9 receptors, arranged in a 3×3 square centered on one receptor. This is illustrated with two sample sets of fibers connecting the receptors to the convolution array. The receptor clusters for the two samples are picked out with shading to help illustrate what each one is "looking at" in the receptor mosaic. Each receptor cluster defines the ***receptive field*** of the associated convolution cell.

[*Terminological note repeated from Ch 3*: The standard definition of receptive field is the patch of retina that influences the output of the cell in question. Strictly speaking, it is not defined in terms of the pattern of weights associated with that area, but in practice the term receptive field is often used to refer to that pattern, as we do in this book. For the latter usage, the receptive field concept is similar to *template, operator, mask,* or *weighting function* used in the computer vision literature.]

The two sheets of cells, receptor mosaic and convolution array, are shown in **5.7** neatly lined up one over the other. That is, the central cell in each receptor cluster feeds a convolution cell whose position in the convolution array has a matching spatial location. This helps us keep track of what is going on in the figure but the physical layout would not be critical in a vision system. The key property is where the connecting fibers come from and go to, a point that was made in Chs 3 and 4 in connection with *place coding*.

For simplicity, only the fibers to three convolution cells from the receptor mozaic are shown in in **5.7**, but in fact every convolution cell would be connected to the receptor mosaic in a similar fashion. The key difference between convolution cells is that the set of inputs defining their receptive fields come from receptor clusters in slightly different positions in the receptor mosaic. There would be hundreds of fibers linking the two arrays. In **5.7** there are only 72×45 fibers but there would be very many more in a realistically sized vision system.

Because of these multiple connections, each receptor has to feed many different convolution cells. Again for reasons of simplicity, **5.7** shows only one instance of this kind—this is the receptor picked out with the arrow+label. In other words, the two convolution cells in question have wiring connections such that they share one receptor in common.

The small inset in **5.7** shows an example of a possible set of ***receptive field weightings,*** which are the same for each cell in the convolution array. For example, if a receptor is signaling 18 units of activity (this number is its code for the associated image intensity) then the influence transmitted to the convolution array is much smaller than this, in fact only 1/9th. Why 1/9th? Because there are 9 inputs in this particular 3×3 receptive field, and so multiplying each one by 1/9th comes to the same thing as adding up all the 9 inputs and dividing by 9 to get the mean. The conclusion is that weighting inputs is a neat trick for doing the averaging arithmetic we require.

[To see why this works, consider if all the 9 receptors in a receptive field had the same activity level, say 18 units. The neighborhood total for all nine will come to 9 × 18 = 162. Obviously, dividing this by 9 to get the mean will give 18, which is the answer we want because all receptors in this example were registering 18 units of activity. Now consider doing the same thing by weighting. Each input fiber delivers 18 × 1/9 = 2 units of activity, and adding up all 9 inputs gives 9 × 2 = 18.]

This is good news because weighting is convenient if neurons are all you have to do the job. They work by transmitting excitation or inhibition, Ch 3. In **5.7** all the fibers from the receptors pass on excitation, with the precise amount of excitation adjusted to suit the weighting required for the computation in question.

To summarize so far: convolution using suitably weighted connections is one way to implement a local neighborhood averaging noise cleaning algorithm. This is said to be ***biologically plausible*** because it embodies a procedure that lends itself to being readily implemented in neurons.

Convolving an input image with a receptive field is a widely used technique for processing images, in both man-made and biological visual systems. It is essential to understand convolution to understand vision.

In the case we have been considering, convolution replaces the original grey level description with a smoothed grey level description, which is said to be the ***convolved image.*** However, many other sorts of receptive fields can be used and for them the convolved image will not be a more or less close replica of the original grey level description. Rather, the activity levels in the convolved image will vary according to the "goodness of fit" of the receptive field at each location in the input image. The "bar-tuned" receptive fields we investigated in Ch 3 are a case in point, and we will see other examples in due course. For the moment, remember that each element in the convolved image has a position which signifies the point in the input image on which the receptive field was centered when the receptive field's goodness of fit with the relevant point of the image was calculated.

To reiterate the key terminology, convolution can be defined as applying a receptive field all over an image.

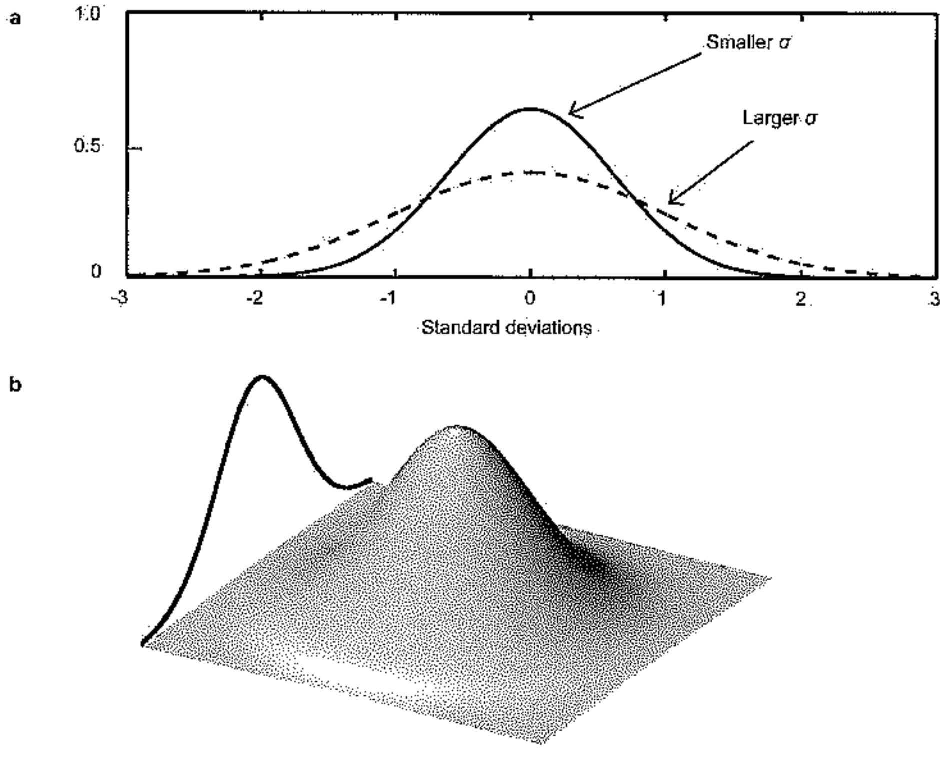
Equivalent Algorithms

Summarizing the simple weighting trick for implementing noise cleaning by taking a local neighborhood average gives us:

Step 1 For each pixel *P*, multiply its intensity and that of its neighbors by 1/*N*, where *N* is the number of elements in the receptive field centered on pixel *P*.

Step 2 Add up the values so obtained, and put the result into the convolution cell corresponding to the pixel *P*.

The point of giving this second procedure in this form is to emphasize that a computational theory of a task can usually be implemented with one of a number of algorithms. Each one ***implements*** the



5.8 Gaussian distributions

a Cross-sections of two gaussians of different widths or "spreads." The parameter that determines the size of the spread is the standard deviation (σ, pronounced sigma). Weight strength is plotted on the vertical axis.

b 3D picture of a two-dimensional gaussian, with its profile shown as the graph on the left.

same theory and achieves an equivalent output but each does the job in a different way. Which one is chosen will be determined in part by the hardware or brainware available for running the procedure. Our interest in biologically plausible procedures leads us naturally to the kind of weighting just set out. We will see weighting of this kind in the receptive fields of the bipolar cells of the retina, in Ch 6.

The noise cleaning receptive field used in **5.7** raises some interesting questions. For example, how big should the receptive field be? Is it sufficient to use in general just the closest 8 neighbors for each pixel to smooth the noise away? Or should a larger size than 3 × 3 be used?

It turns out that biological vision systems have a range of receptive field sizes. Before explaining why, we need to delve a bit deeper into the question of now best to average out noise.

Gaussian Smoothing

We said earlier that the 3 × 3 averaging procedure was a very simple scheme for noise cleaning, chosen to introduce some core ideas. It turns out on closer inspection that instead of all the pixels having equal weights when feeding into the convolution array those near to the center of the receptive field should be weighted more highly and those further way less so.

The particular shape of the weighting distribution has to optimize two conflicting goals: smoothing away noise and not disturbing too greatly where edges are to be found in the convolved image. There is a theorem (see Marr and Hildreth, 1980) stating that the shape of weighting distribution that achieves an optimal trade-off between the two goals is bell-shaped. Examples are shown in **5.8a**.

The technical name for a distribution with this shape is a ***gaussian***, after the mathematical genius Karl Friedrich Gauss (1777–1855) who discovered it. As can be seen in **5.8a,** the pixels nearest the center of the receptive field have largest weights and thus have most influence. The weights then reduce, smoothly and gradually, to zero at the boundary of the field. This pattern of weights is shown in **5.8b** as a three-dimensional view of a receptive field, with weight strength again plotted on the vertical axis.

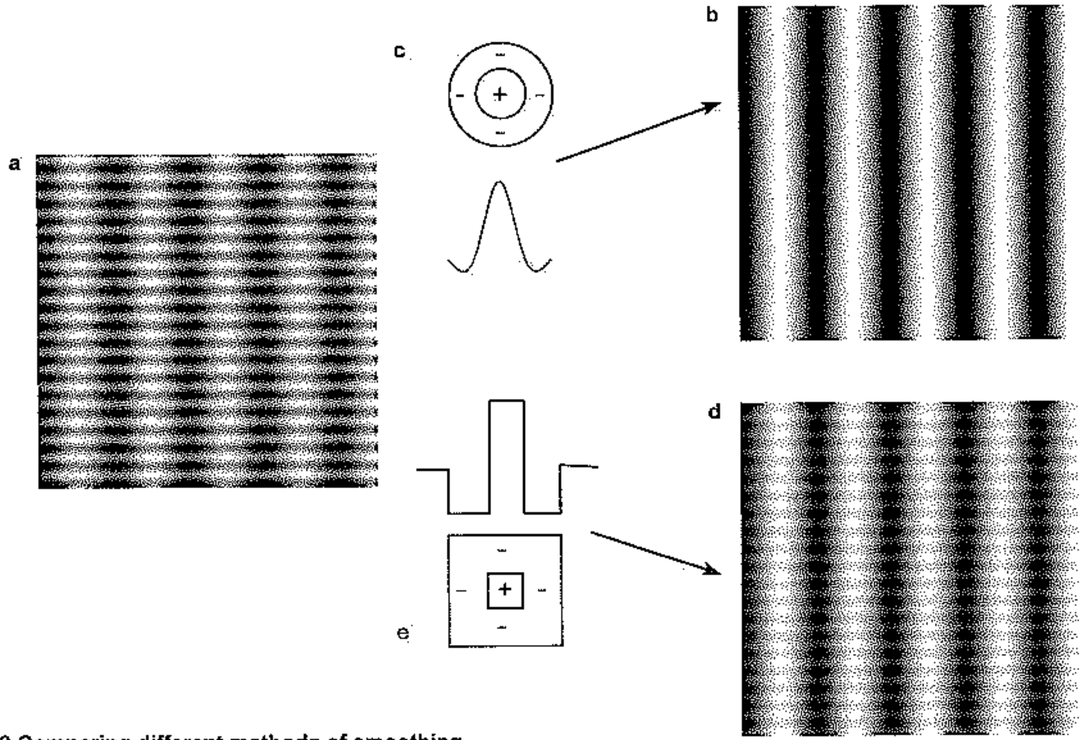
An example of the benefit of gaussian smoothing over smoothing with a rectangularly shaped field is shown in **5.9.** The former but not the latter uncovers the vertical grating to which a weaker higher spatial frequency (Ch 4) horizontal grating has been added as noise.

Image Edges have Different Scales

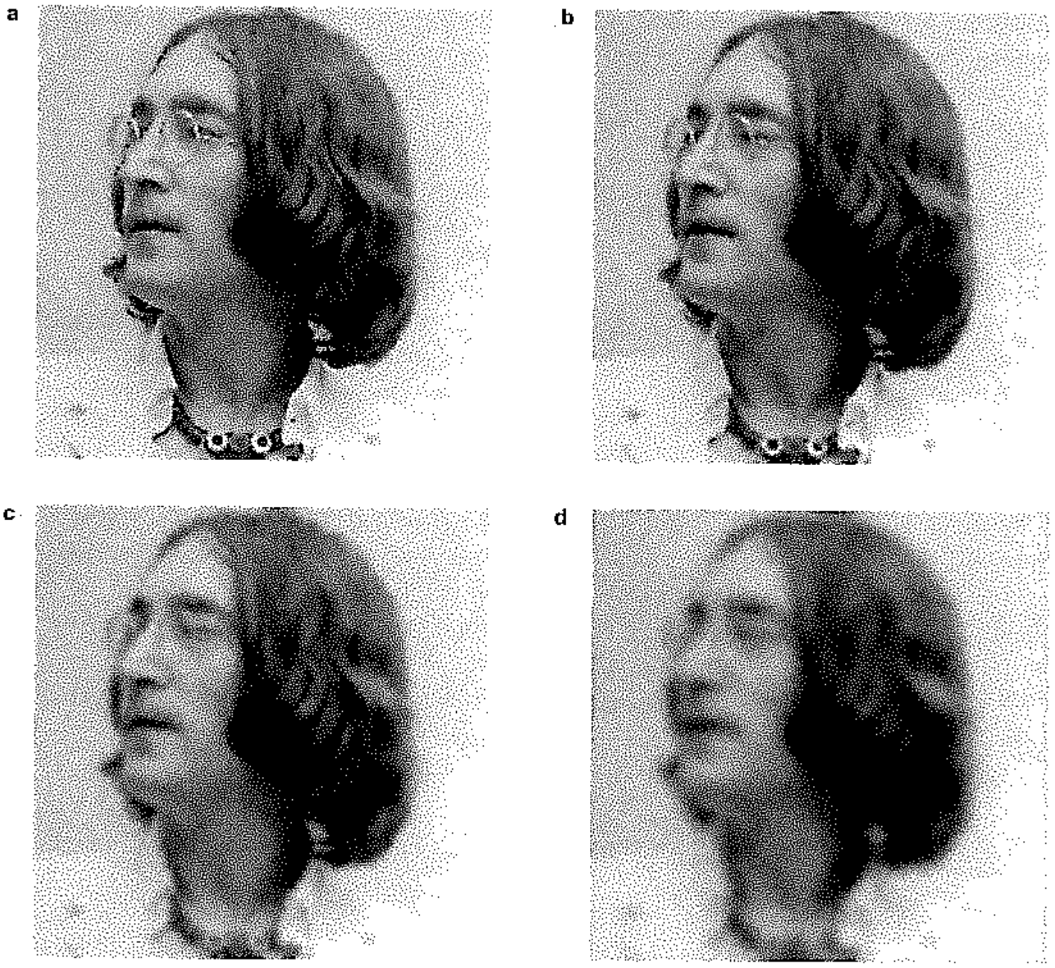
A close look at the image intensity landscape of the John Lennon image in **5.4** reveals that its variations in intensity range over different ***spatial scales.*** That is, some intensity changes occur over small image regions and are thus quite sudden, e.g., the steep, sharp edges from the spectacles. Others occur more gradually over larger image regions, e.g., the shallow, fuzzy edges from shadows around the chin. The profiles of two slices of this image shown in **5.6** nicely illustrate this fact.

The different scales reflect different sorts of scene entities. Sharp edges often arise from object boundaries or texture markings, whereas shallow ones usually arise from shadows. It is important to find a way of dealing with this fundamental fact of scale in developing a theory of edge detection. (Small and large scales of image intensity changes are often referred to as conveying high and low ***spatial frequencies*** respectively, for reasons made clear in Ch 4.)

One way of finding edges at different scales is to use a range of receptive field types, each one "tuned" to a limited range of edge scales. Each



5.9 Comparing different methods of smoothing
a Vertical grating to which has been added a horizontal grating of equal contrast.
b Smoothing a with a circular symmetric gaussian filter with the receptive field of the kind shown in plan view and in profile in **c**. This preferentially attenuates the narrow (horizontal) grating.
d Smoothing a with a rectangular-shaped filter with a profile of the kind shown in **e**. This rectangular filter does not attenuate the narrow (horizontal) grating as much as the gaussian filter.



5.10 Convolving the Lennon image with gaussians of different standard deviations, σ
From **a** to **d**: σ = 1, 2, 4, and 8 pixels of the receptive field used in the convolution.

such receptive field is said to operate as a ***filter*** because it "lets through" only edges at a particular scale. But what are the principles for designing scale-tuned receptive field filters? This question, as usual, forces us to find a decent computational theory of the task.

Finding Edges at Different Scales

It was David Marr and Ellen Hildreth who first clearly stated, in 1980, edge detection as an optimization problem. Specifically, they stated that edge detection required an optimal trade-off between two conflicting tasks: selecting edges at a given scale and accurately localizing the positions of edges in the image at that scale. This led them to use gaussian receptive fields spanning a range of sizes, thereby exploiting the theorem referred to above in connection with noise cleaning. This is a nice example of a computational theory, because it is a mathematically well-founded solution to a clearly stated problem.

So, filtering for different scales is achieved by having a range of different convolution arrays, each one using a differently sized receptive field, but all having weightings of the basic bell shape, **5.8**. The ***parameter*** that determines the size, or equivalently the spread, of a gaussian is its ***standard deviation***, ⊠ (pronounced sigma). Profiles of gaussians with large and small ⊠ are illustrated in **5.8a**.

Results from using the Marr/Hildreth scale filtering scheme on the Lennon image are shown in **5.10** using four gaussians, each with a different ⊠.

As can be seen, the smallest (⊠=1 pixel) preserves the sharpest image intensity variations. The image is increasingly blurred as ⊠ is enlarged—compare the results for ⊠=1, 2, and 4 pixels. The largest bell-shaped receptive field completely removes the fine variations in intensity, both those due to noise and to sharp scene edges.

Computational Theory for the Task of Measuring Image Gradients

We started out by saying that image intensities can be thought of as a hilly landscape, **5.4**. This makes it natural to think of edge detection as the task of measuring ***image intensity gradients***. So we now need a task theory telling us how to measure these gradients in our scale-filtered images. A simplified account of this is possible using the metaphor of the input image intensities forming a hilly landscape because the task then becomes one of measuring the gradients of hills. Technically, this is called finding the ***first derivative*** of the image landscape.

A road sign saying 14%, or 1 in 7, warns you that you will shortly go up or down 1 meter for every 7 meters travelled along the road. The numbers 1/7 = 0.14, or 14% , are called ***gradients***. This example shows that measuring a gradient means working out a ratio: the change of height for a given horizontal extent. In our case, this translates to finding edges by measuring how much intensity goes up or down over a given region of the image. That is as much gradient measuring theory as we need in this introductory account.

Algorithm for Measuring Image Gradients

The task theory tells us that we should measure image gradients and what this means. We now need a biologically plausible procedure for doing it. A simple algorithm is:

Step 1 Measure the difference in intensity (height) between image points.

Step 2 Divide this difference by the distance between the points, to obtain the gradient.

As an example, consider a case of the image intensity gradient between neighboring pixels, whose intensities happen to be 10 and 5 units. The gradient is then simply (10 − 5)/1 = 5 units of intensity per unit of distance. The latter is measured in pixel widths, which in this example is 1 as it concerns neighboring pixels.

Can this be procedure be realized in neurons? An equivalent but more suitable scheme turns out to be:

Step 1 For each pair of neighboring pixels, weight the inputs from one as positive, the other as negative, by multiplying by +1 and −1 respectively.

Step 2 Add together the weighted inputs.

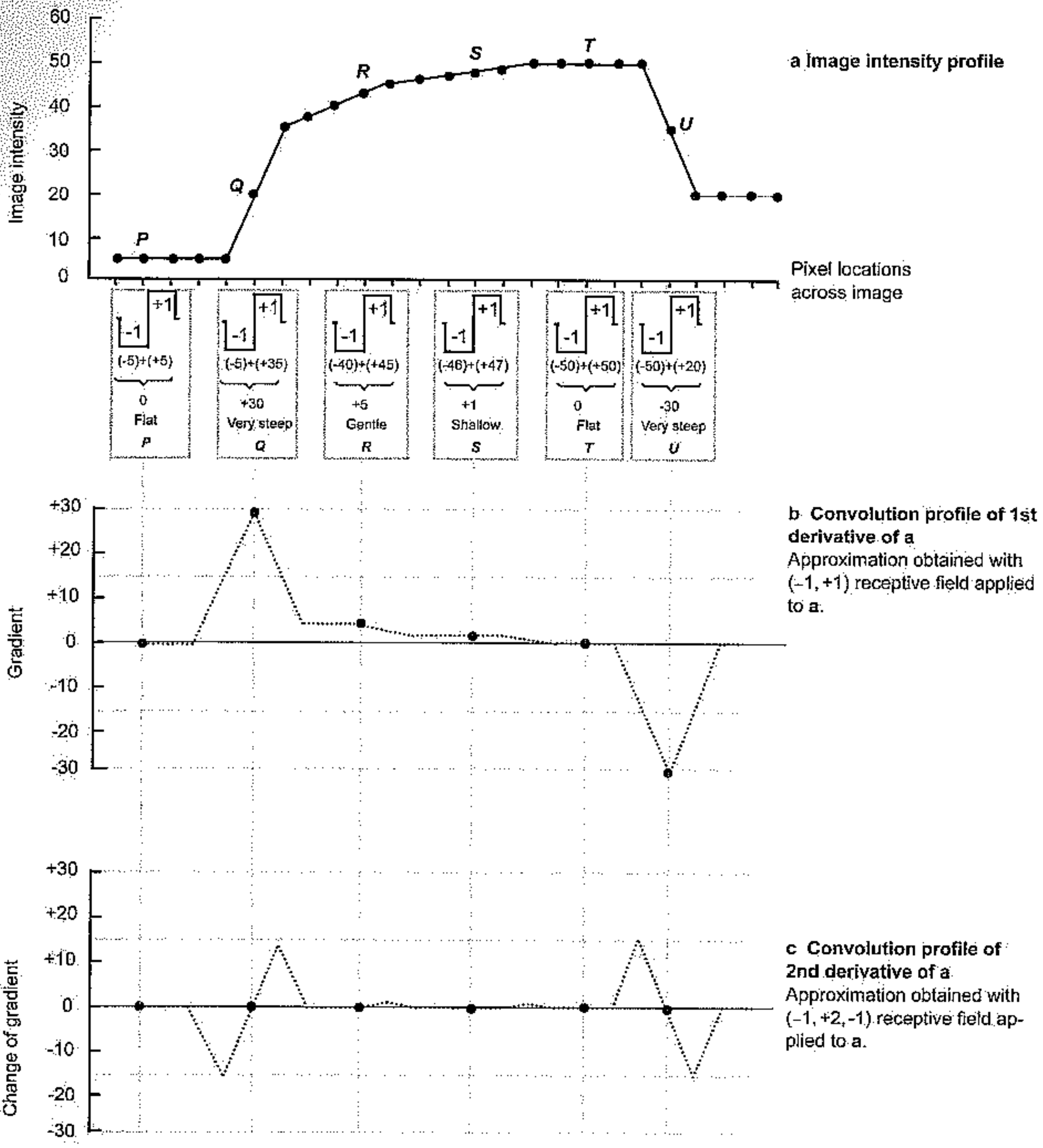
This procedure is arithmetically equivalent to the previous one. It is biologically plausible because it is easy to implement with neurons.

Measuring Gradients in a Slice of an Image Intensities Landscape

To illustrate this procedure it is helpful to begin with just a single slice of an image intensity landscape, and then return later to considering the whole image. The slices shown in **5.6** are a bit complicated for showing the basic ideas and so an artificially generated slice is illustrated in **5.11**. It shows a cross-section in which, working from the left side, intensities rise and then fall, reflecting the presence of a bright ridge. Note that the gradient at *Q* is steeper than at *S*: walking up the hill at *Q* increases your height considerably more than walking at *S* for the same lateral distance.

The weighting of neighboring pixels in **5.11** is achieved by using a receptive field with −1 and +1 weights. For example, the pixel value on one side of the point labeled *P* is multiplied by −1 and the pixel value on the other side by +1, and then the two quantities so obtained are added together. In the case of *P*, this gives (−5) + (+5), which equals 0. Zero makes sense in this case as *P* is on a horizontal part of the image intensity profile. The zero result is entered in the graph labeled convolution profile, **5.11b**.

The same procedure is also applied in **5.11** for the points *Q*, *R*, *S*, *T*, and *U*. But of course, as the gradient of the whole profile is required, it is necessary to apply the (−1, +1) receptive field all along the image intensity profile shown in **5.11a**, rather than just at the chosen points *Q*, *R*, *S*, *T*, and *U*. Applying a receptive field all over an image is called convolution, as stated above. The image is said to have been *convolved with* the receptive field.



a Image intensity profile

b Convolution profile of 1st derivative of a
Approximation obtained with (−1, +1) receptive field applied to a.

c Convolution profile of 2nd derivative of a
Approximation obtained with (−1, +2, −1) receptive field applied to a.

5.11 Measuring gradients and changes of gradients using receptive fields
The units of intensity plotted in a are arbitrary but chosen to convey the main ideas.

To sum up so far, **5.11b** is the convolution profile derived from **5.11a**, using the (−1, +1) receptive field. This profile illustrates what is called the *first derivative*. It shows that when the intensity profile is flat, the output given by this receptive field is 0, as required. On the steepest parts in this example, the output is +30 or −30 (which means 30 units of intensity change for a shift across the image of 1 pixel). The plus sign indicates an up gradient, and the minus sign indicates a down gradient. The slope of the shallowest non-zero gradient is +1, e.g., at point *S* in **5.11a**.

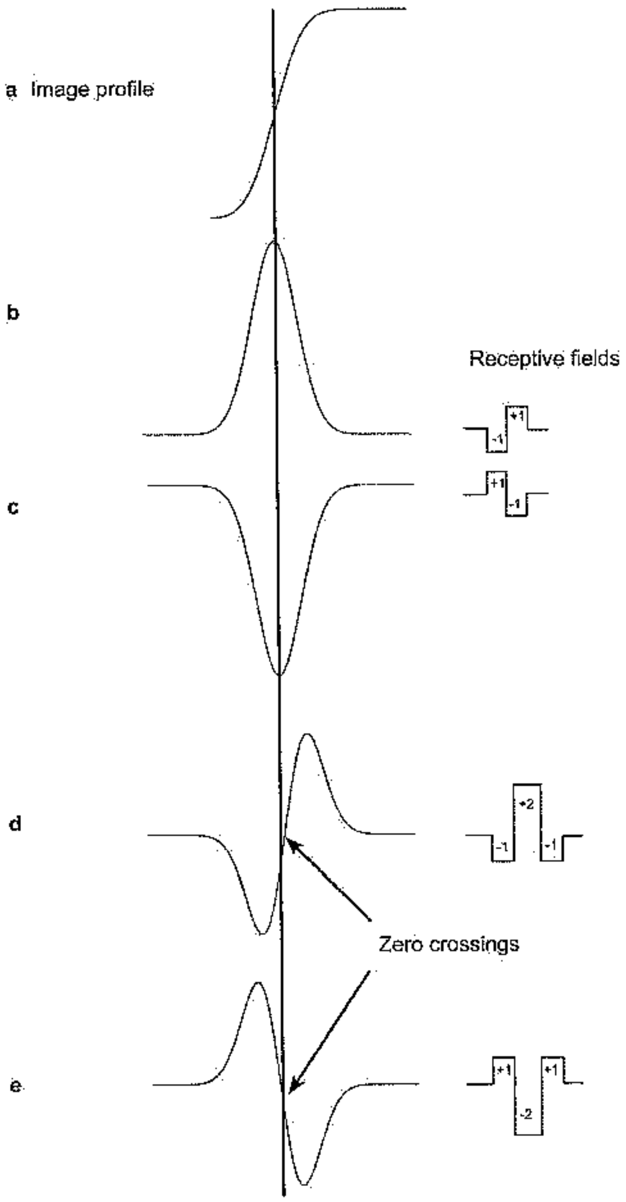
Measuring Changes in Gradients

Measuring gradients (finding the first derivative) is a useful thing to do but it is also desirable to extract information about *changes in gradients*. Doing this is called finding the ***second derivative***. It may seem a bit odd at first to think of measuring changes in changes. However, changes in intensity gradients are significant because they are usually caused by things in the world that the visual system wants to know about, such as illumination changes, surface orientation changes, and changes in surface reflectance (e.g., the edges of objects or surface markings).

One way in which a gradient change can be located is to search for a peak in the first derivative profile. The changes in slope in **5.11a** either side of *Q* produce a sharp peak in its first derivative, **5.11b**. Note that a trough (a "negative peak') occurs for the point *U*. This point is similar to *Q* in that it has sharp changes of slope on either side. However, a trough occurs in the convolution output rather than a peak simply due to the signs of the (−1, +1) receptive field coupled with the downward direction of the slope at *U*. If a (+1, −1) receptive field had been used then the trough at *U* would have been a peak, and the peak associated with *Q* would have become a trough.

To summarize, the first derivative measures gradients and the second derivative measures changes in gradients. The latter can be found by measuring gradients in the first derivative.

The second derivative of the intensity profile in **5.11a** is shown in **5.11c**. Note that at points where a peak or trough is located in the first derivative, **5.11b**, the second derivative passes from positive to negative. Such a point is called, natural-



5.12 Convolving receptive fields with an edge
The intensity change in **a** gives rise to a peak or a trough in the first derivatives **b** and **c**. It gives rise to a zero crossing in its second derivatives **d** and **e**.

ly enough, a ***zero crossing*, 5.12**. Hence, there are two ways of locating changes in intensity gradient: either find zero crossings in the second derivative or find peaks or troughs in the first derivative.

Single-Step Way of Measuring Changes in Gradients

We have said that the second derivative can be found by applying the (−1, +1) receptive field twice over. That is, use the following algorithm:

Step 1 Measure gradients using receptive fields of the type (−1, +1) or (+1, −1).

Step 2 Measure gradients in the output of Step 1, again using fields of either (−1, +1) or (+1, −1).

It turns out that there is a way of measuring changes in gradients using a single processing step. The receptive fields for doing this have profiles of either (−1, +2, −1) or (+1, −2, +1). We now explain why this works.

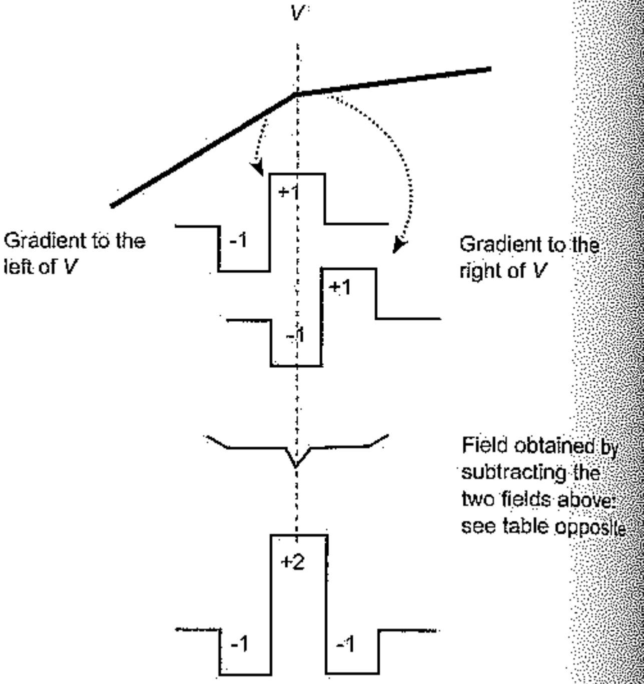
At a point where the gradient in the image intensity profile changes, the gradient to the right of the point is different from the gradient to the left. This is illustrated in **5.13** for a point labeled *V*. So if we measure the gradient to the left and subtract from it the gradient to the right we obtain a measure of their difference and hence a measure of the second derivative.

An easy way to do this in a single-step convolution is to use a receptive field built up in the following way:

	V		Point *V* is where the 2nd derivative is being measured
−1	+1		Gradient to the left of *V*
	−1	+1	Gradient to the right of *V* for subtraction
−1	+2	−1	Resulting receptive field

Beware a possible confusion at this point. In the middle column of the table the subtraction can be written as +1 minus (−1) = +2. Recollect that a rule of arithmetic is that "two minuses give a plus." That is, doing the minus operation twice over yields a plus. Thus, minus (−1) becomes +1, and this is why the answer is +2.

The result of applying this receptive field to the intensity profile in **5.11a** is shown in **5.11c**, in which the zero crossings mark points of gradient change in the image. We will have more to say about zero crossings as edge location markers in Ch 6.


V
Gradient to the left of V
Gradient to the right of V
+1
-1
+1
-1
Field obtained by subtracting the two fields above: see table opposite
+2
-1
-1

5.13 Receptive field of (−1, +2, −1) type measures changes of gradient
This figure should be studied in conjunction with the associated table shown opposite.

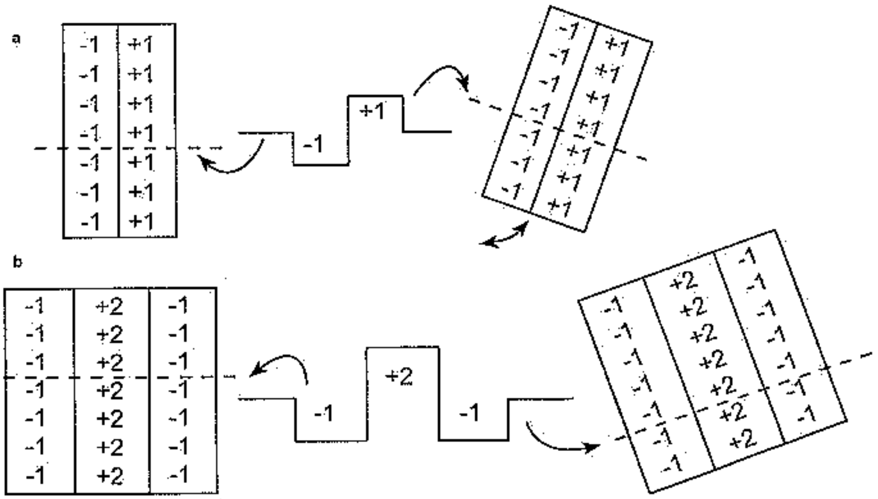
The same simple arithmetic can be used to derive the (+1, −2, +1) weighting profile by subtracting neighboring pairs of (+1, −1) fields.

The conclusion we have thus come to is: convolving with a (−1, +2, −1) or (+1, −2, +1) receptive field is a neat way of getting the second derivative in a single processing step. This could explain why many cells in biological vision systems have these sorts of fields.

Two-Dimensional Convolutions

For purposes of exposition, only a one–dimensional slice of an image intensity profile has been considered so far. However, a visual image is usually a two–dimensional array of intensities. So how can the receptive field profiles just described can be extended to cope with two dimensional images?

A straightforward way to do this is to keep the receptive field functionally a 1D device by making it sensitive to gradients in one particular orienta-



5.14 Oriented receptive fields for measuring directional derivatives
Fields with different orientations are shown together with the profiles across them in the directions shown with the dashed lines that are perpendicular to field orientation. **a** 1st derivatives, **b** 2nd derivatives.
Note the similarity of these fields to those of the simple cells described in Ch 3.

tion only. This is done by the simple expedient of elongating each receptive field, **5.14**. In this case the operators are said to provide ***directional derivatives*** because the derivative is tied to a particular orientation.

Having read Ch 3, you may recognize how these extended one–dimensional receptive fields bear a strong resemblance to the receptive fields of some of the *simple cells* found in the striate cortex of cats and monkeys by Hubel and Wiesel. This similarity makes it tempting to suggest that simple cells really are devices for delivering directional derivatives. If this is so, gradient measurements would need to be sampled in a fairly large number of different orientations at each point in the image. This may not be a great handicap for the brain because it has lots of orientation tuned cells (more details on these in Ch 9).

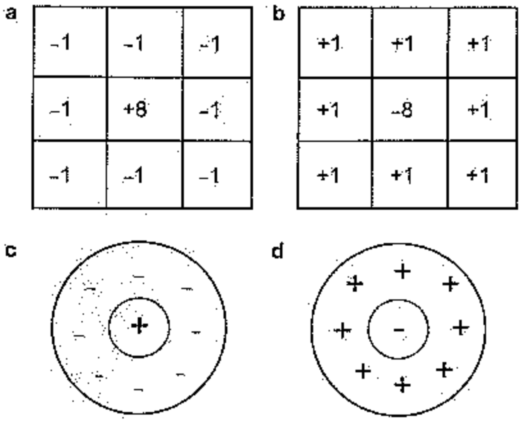
The idea that the receptive fields of simple cells can be interpreted as providing oriented second derivatives shows how far we have come from Ch 3. There we discussed simple cells as candidates for "bar detectors." Careful examination of the task of edge detection has hugely refined our understanding of what these cells might be doing. This is a fine example of the benefits that can come from task analysis at the computational theory level.

That said, it is also important to note that whether simple cells really are best thought of as delivering directional derivatives is still an open question. We know too little about brain mechanisms to be sure at present. Even so, some believe it to be the current "best bet" for the functional role of at least some types of simple cells.

Measuring the Image Gradients Using an Isotropic Receptive Field

Is it possible to measure the first and second derivatives (i.e., image intensity gradients and changes in image gradients, respectively) in two–dimensional images *without* using oriented receptive fields? That is, can two-dimensional receptive fields be constructed for measuring derivatives that are sensitive to image intensity changes irrespective of their orientations? The technical term for describing any process that is the same in all directions is ***isotropic*** (*iso* = equal; *tropia* = direction).

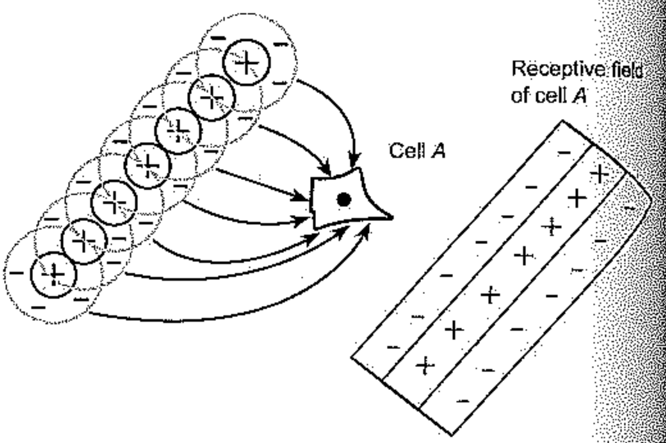
It turns out that the first derivative cannot be measured with an isotropic receptive field but this can be done for the second derivative using a ***laplacian*** receptive field, **5.15a.** You can think of this receptive field as being created by spinning a (–1, +2, –1) set of weights around its center. We will see in Ch 6 that this type of receptive field is similar to those possessed by the certain cells in the retina, **5.16.** This will prove important when we consider what certain retinal cells seem to be doing and how they feed into brain cells.



-1	-1	-1
-1	+8	-1
-1	-1	-1

+1	+1	+1
+1	-8	+1
+1	+1	+1

5.15 Laplacian receptive field
a A 3x3 array of weights that can be thought of as a combination of four (−1, +2, −1) fields, one each for vertical, horizontal and the two obliques directions.
b A similar combination of four (+1, −2, +1) fields, producing a receptive field with opposite signs to those shown in **a**.
c and **d** Respectively, on-center and off-center circularly symmetric receptive fields of the antagonistic center-surround kind found in many biological visual systems. Their qualitative resemblance to the laplacian fields in **a** and **b** is striking. We discuss in detail receptive fields of this type in Ch 6, on the retina.

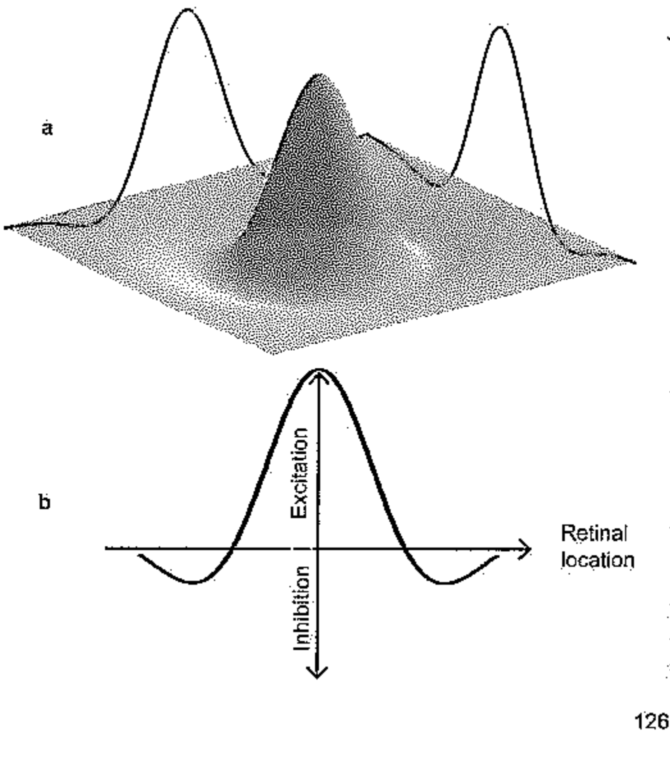


5.17 Combining a set of isotropic gradient measurements to create a cell signalling an orientation-tuned second derivative
The circularly symmetric fields measure changes of gradient in slightly different positions on the image. When their outputs are fed into *cell A* then this cell has a oriented receptive field of the kind possessed by some simple cells. *cell A* can thus be regarded as an *operator* implementing a ***figural grouping process*** that links edge points: see Ch 7 for details. This is different way of interpreting what simple cells may be doing.

It is easy to compute directional derivatives from isotropic gradient measurements. This is done simply by combining the outputs of circularly symmetric (isotropic) receptive fields dealing with suitably located nearby image points. An example is shown in **5.17**. One way of looking at this is as a *figural grouping operation* which links together edge points that form an edge with a certain orientation (details in Ch 7).

Skip from Here to Summary Remarks?

Readers who do not want technical details on how to combine blurring with measuring image gradients can move on to the Summary of this chapter on p. 132, and then on to Ch 6.



5.16 Circularly symmetric receptive field
This field is responsive to edges of all orientations, as demonstrated in detail in Ch 6.
a 3D picture of the response profile of an on-center receptive field. Fields of this kind are informally called *Mexican hat* receptive fields because of their similarity to the cross-sectional shape of a classic form of Mexican head wear.
b Explanation of the profiles shown in **a**. The graph rises above the horizontal "zero response" line when excitatory influences arising from light falling on to the field's center exceeds inhibitory ones arising from light falling on the surround.