Johnson-Laird (1988) The computer and the mind.

CHAPTER 4

The visual image

Consider three beliefs about vision. The first, which is often expressed, is that it is trivially easy. The eye is like a television camera. You point it at a particular scene, it registers that scene, and projects a picture of it inside your head.

The problem with this tale from 'folk psychology' is simple: who looks at the picture, and how do they see it? Clearly, the end result of perception cannot be a picture, because that in turn would have to be perceived, just as a picture hanging on a gallery wall makes no sense until it has been perceived. The explanatory problem has merely been thrust backwards a single step. In computational terms, this way of regarding vision is useless because it fails to deliver the right result.

The second belief is that vision is impossible. The eye is like a television camera. You point it at a particular scene, and it registers that scene. But many different arrangements of things can produce the same image, and so the brain is unable to work out which particular arrangement you are looking at.

This more sophisticated argument can be illustrated by an example. Suppose that the eye views a single thin bar that appears upright. In fact, the bar could be leaning towards the viewer or leaning away and yet project virtually the same image on the retina. Figure 4.1 illustrates three possibilities: each of the bars creates the same image since from top to bottom they all project the same visual angle. Although the figure contains only three possibilities, the eye would receive the same image from an infinite number of different bars provided that their lengths, thicknesses and positions combine to produce the appropriate visual angle.

A natural response to this sceptical argument is to say: I have two eyes, they receive slightly different images of the bar, and I can use this disparity to work out its true orientation. Similarly, if

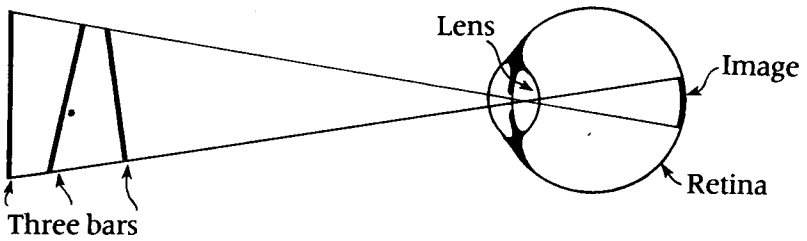


Figure 4.1: Three different bars that project the same image on the retina.

I walk around, or move my head from side to side, I will see the bar from different viewpoints, and the resulting images will be consistent with only one particular orientation. To which the sceptic replies: very well, perception is impossible if you close one eye and stand motionless; and, by the way, you owe me explanations of how the brain does stereopsis, that is, how it recovers depth from the slight disparities between the two retinal images, and of how it combines images from different viewpoints.

The third and most helpful view is that vision is easy for the brain to do, but difficult for us to understand. You normally see things automatically and with no effort at all. This ease yields the subjective experience of direct contact with the world, which has an enormous evolutionary advantage. If you see a tiger, you avoid it. You do not stop to inspect the processes of vision to check whether they are working properly. What is advantageous to the species, however, creates a problem for cognitive science. It is hard to discover how vision works.

To solve any such problem, we must bear in mind a lesson from computation: we need at least three different levels of explanation. We need a theory of *what* is computed – what the input to the process is, what has to be recovered from it, and what constraints may guide the process. We need a theory of *how* the system carries out the computations, that is, a computable theory of the procedure that it uses. We need a theory of the underlying neurophysiology (the 'hardware' of the nerve cells in which the procedure is embodied). Investigations that are restricted to only one of these levels are unlikely to yield a correct explanation. A theory of vision that embraces them all has been framed by the late David Marr and his colleagues, particularly Keith Nishihara, Tomaso Poggio and Shimon Ullman. It is their account that I shall mainly follow, because they have modelled it in computer programs that will be useful for the robot's vision.

THE COMPUTATIONAL PROBLEM OF VISION

The hapless robot, which we left in the previous chapter wandering about its universe, finds its way home solely by 'dead reckoning'. Vision would enable it to see where it is going, but for that it will need eyes. The function of an eye is twofold. It must contain an array of light-sensitive elements – a retina – to convert the energy in quanta of light into an internal symbolic code. Light is reflected by surfaces in all directions, and so the eye must also ensure that the light from each point in a scene falls on to a single point of the retina, not on to all of them. A small pin-hole, or a lens, focuses the light in this way. That's the theory, but alas, the human eye, as Helmholtz remarked, is not a perfect instrument.

The illumination, arrangement of the surfaces in the scene, and the amount of light they reflect, cause a particular pattern of light to fall on to the retina. The cells in the retina convert this energy into nerve impulses. These nerve impulses depend on the physics of light and lenses and on the biochemistry of the nerve cells in the eye. Cognition is therefore not wholly a matter of computations that transform mental symbols: symbols can be created by physical interactions with the world. These interactions are hardly mysterious. Electronic cameras, which the robot will use, depend on sensors to convert the energy in light into electrical impulses, albeit in a much cruder way than the human eye. We shall not be concerned with these details, but with what happens afterwards.

For the robot, what is ultimately to be computed is a symbolic representation of the three-dimensional world that can guide its behaviour. As it moves its eyes over a scene, the representation will make *explicit* where the robot is in relation to the objects in

the scene, and where they are in relation to one another. (The reader will recall from Chapter 2 that a symbol makes information explicit to a process if that information is available without further computation.) The representation will also make explicit shapes, colours, textures and illumination. If the robot moves through the world, then the representation will be kinematic, indicating that the robot is moving (and the scenery is not). The robot will be able to use the representation to navigate its way around objects without bumping into them or falling down holes. If the things in the scene are moving, then the robot will be able to use its representation to predict their future positions. If two things collide, then the representation will be dynamic: the robot will see that one thing *causes* another to move. If things are familiar to the robot, it will recognize them for what they are. It will even be able to identify objects that it has never seen before, such as a tree of an unfamiliar sort or a bookcase of an unconventional shape. In sum, the robot's vision, like that of a human being, will construct a *model* of the world from the patterns of light falling on its retinae.

CONSTRAINTS ON THE VISUAL PROCESS

Empiricists think of the mind as a blank slate, and of the information that makes vision possible as out there in the real world. Rationalists think of the mind as containing knowledge that it imposes on the world. The truth appears to be that both are right. There is information in the patterns of light that fall on the eyes, but visual processes also depend on assumptions about the nature of the physical world. Vision is rather like the problem of finding the value of x in the equation:

$$5 = x + y$$

The equation is plainly not well posed, since there is no way of determining how much of the 5 comes from x and how much from y. If, from prior knowledge, you can assume that y is likely to have a value less than 1, then you can infer that x is likely to have a value greater than 4. In vision, assumptions about the world make an ill-posed problem soluble.

The late J. J. Gibson emphasized that light reflected from surfaces and focused on the retinae contains a large amount of information. The proposition has been demonstrated by an analysis of the projective geometry of two images. Christopher Longuet-Higgins has shown, amongst a number of such proofs, that if just five points on the surface of an object can be matched from one photograph to another taken from a different angle, then the complete three-dimensional geometry of the situation can be established – the orientation of the surface with respect to the two cameras, and the relative positions of the cameras to each other. Berthold Horn and his colleagues have constructed analogous proofs about shapes: the shape of something can be recovered from the intensity values of its image provided that its surface is smooth, matte so that it reflects light in a uniform and diffuse way, and has a known orientation at certain points in the image.

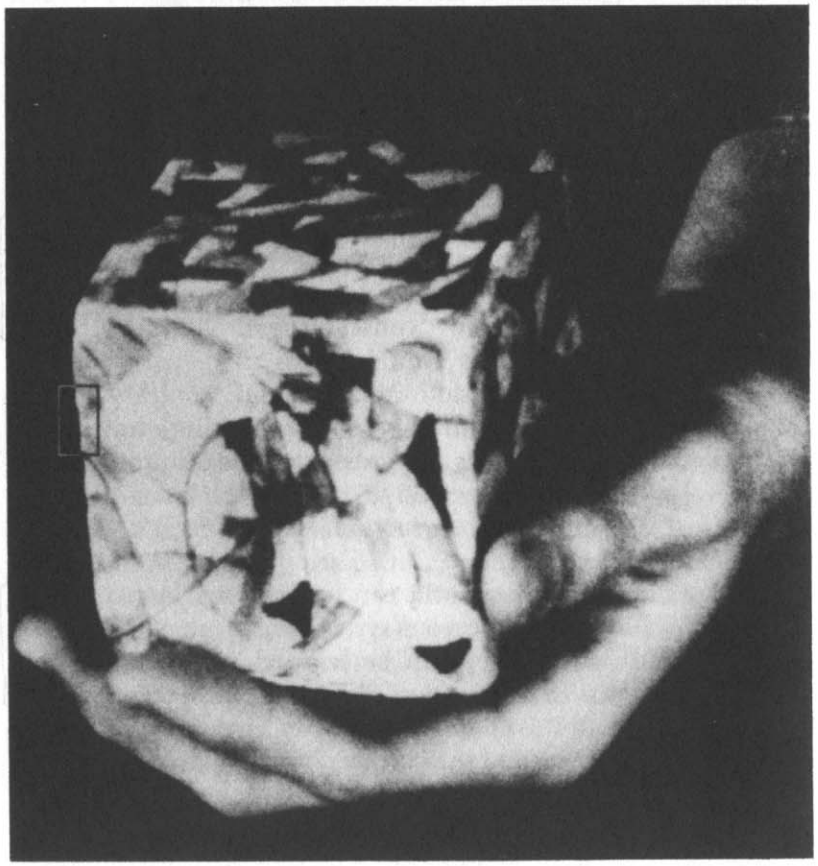
No matter how much information is in the light falling on the retinae, there must be mental mechanisms for recovering the identities of things in a scene and those of their properties that vision makes explicit to consciousness. Without such mechanisms, retinal images would be no more use than the images produced by television cameras, and, contrary to the naive view, *they* cannot see anything. If the robot is to travel over a dangerous terrain, it needs to process the images produced by its cameras so as to recognize and to locate obstacles and pitfalls. These processes must rely on certain assumptions about the world.

Assumptions about the world can be acquired in two ways. They can be built into the nervous system as a result of millions of years of evolution, or they can be learned during a person's lifetime. Both sorts of assumption play a part in perception, but in different ways. In this chapter, I shall concentrate on the initial stages of vision that lead to the construction of a visual image; in the next chapter, I will describe the perception of depth. Both depend on general constraints that are inborn. In Chapter 6, I will turn to knowledge acquired in a lifetime and how it enters perception.

THE FIRST STAGE OF VISION: THE GREY-LEVEL ARRAY

The first stage of vision is the physical interaction between light focused on the retina and the visual pigment in retinal cells. The human retina is composed of over 100 million light-sensitive cells. A small proportion of them near the centre (about 6

Figure 4.2: A grey-level image: a point-by-point representation of the intensity values of a scene (576 × 454 pixels).



(From T. Poggio, Vision by man and machine. *Scientific American*, April 1984, p. 70)

million) are specialized to respond maximally to light from one of three wavelengths, and these three sorts of cells – the so-called 'cones' – are maximally sensitive to red, green and blue respectively. The full range of perceived colours is based purely on these inputs. The remaining cells in the retina – the so-called 'rods' – are less specialized, responding to light over a wide range of frequencies.

The responses of retinal cells and of an electronic camera comprise, in effect, a two-dimensional array of the intensity values at

Figure 4.3: The intensity values for the part of Figure 4.2 inside the rectangle on the right-hand side.

225	221	216	219	219	214	207	218	219	220	207	155	136	135	130	131	125
213	206	213	223	208	217	223	221	223	216	195	156	141	130	128	138	123
206	217	210	216	224	223	228	230	234	216	207	157	136	132	137	130	128
211	213	221	223	220	222	237	216	219	220	176	149	137	132	125	136	121
216	210	231	227	224	228	231	210	195	227	181	141	131	133	131	124	122
223	229	218	230	228	214	213	209	198	224	161	140	133	127	133	122	133
220	219	224	220	219	215	215	206	206	221	159	143	133	131	129	127	127
221	215	211	214	220	218	221	212	218	204	148	141	131	130	128	129	118
214	211	211	218	214	220	226	216	223	209	143	141	141	124	121	132	125
211	208	223	213	216	226	231	230	241	199	153	141	136	125	131	125	136
200	224	219	215	217	224	232	241	240	211	150	139	128	132	129	124	132
204	206	208	205	233	241	241	252	242	192	151	141	133	130	127	129	129
200	205	201	216	232	248	255	246	231	210	149	141	132	126	134	128	139
191	194	209	238	245	255	249	235	238	197	146	139	130	132	129	132	123
189	199	200	227	239	237	235	236	247	192	145	142	124	133	125	138	128
198	196	209	211	210	215	236	240	232	177	142	137	135	124	129	132	128
198	203	205	208	211	224	226	240	210	160	139	132	129	130	122	124	131
216	209	214	220	210	231	245	219	169	143	148	129	128	136	124	128	123
211	210	217	218	214	227	244	221	162	140	139	129	133	131	122	126	128
215	210	216	216	209	220	248	200	156	139	131	129	139	128	123	130	128
219	220	211	208	205	209	240	217	154	141	127	130	124	142	134	128	129
229	224	212	214	220	229	234	208	151	145	128	128	142	122	126	132	124
252	224	222	224	233	244	228	213	143	141	135	128	131	129	128	124	131
255	235	230	249	253	240	228	193	147	139	132	128	136	125	125	128	119
250	245	238	245	246	235	235	190	139	136	134	135	126	130	126	137	132
240	238	233	232	235	255	246	168	156	141	129	127	136	134	135	130	126
241	242	225	219	225	255	255	183	139	141	126	139	128	137	128	128	130
234	218	221	217	211	252	242	166	144	139	132	130	128	129	127	121	132
231	221	219	214	218	225	238	171	145	141	124	134	131	134	131	126	131
228	212	214	214	213	208	209	159	134	136	139	134	126	127	127	124	122
219	213	215	215	205	215	222	161	135	141	128	129	131	128	125	128	127

(From T. Poggio, Vision by man and machine. *Scientific American*, April 1984, p. 70)

each point on the light-sensitive surface. These values can be represented as numerals (the larger the numeral, the more intense the light). If we ignore colour, these numerals can be converted back into shades of grey and thus into a picture known as the 'grey-level image'. Figure 4.2 presents a grey-level image from an electronic camera, and Figure 4.3 presents the numerals corresponding to the intensities from a small area of this grey-level array. The human retina yields an array with vastly more elements (or 'pixels', which stands for 'picture elements').

The grey-level image is a long way from a representation of what is in a scene. All that it makes explicit is the intensity of light at each point in the array, relative to some arbitrary scale. The next stage of processing recovers more useful information.

THE SECOND STAGE OF VISION: LOCATING CHANGES IN INTENSITY

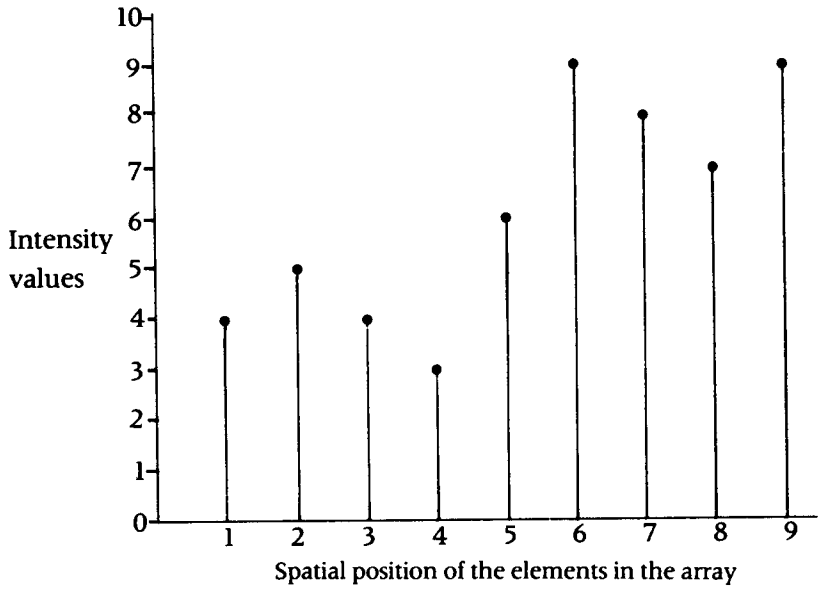
If you squint at the scene in front of you through half-closed eyes, you will see that it is made up of regions of different light intensity. There are bright patches and dull patches, depending on the direction of the light and on how much of it is reflected into your eyes. The intensity tends to change at the edges of objects, and such edges – as the effectiveness of line drawings demonstrates – are important cues to perception. Hence, the visual system must analyse the grey-level array to determine where the boundaries between regions of different intensity occur. However, many boundaries arise from changes in illumination and reflectance rather than from edges, and conversely some edges do not produce clear intensity boundaries. The visual system has to determine which boundaries correspond to the edges of objects.

The information in a grey-level array, such as Figure 4.3, contains a certain amount of 'noise' – random fluctuations occurring in the light itself and within the eye. A simple technique for reducing noise is to replace each value in the array by its local average, that is, the average of it and its neighbouring values. The idea can be illustrated on a small one-dimensional

strip taken from a grey-level array (with deliberately simplified values), such as:

4 5 4 3 6 9 8 7 9

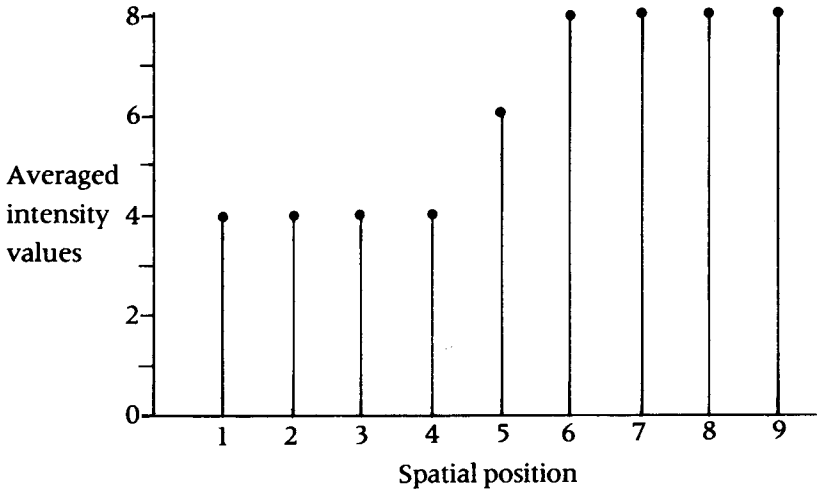
The strip corresponds to the following graph of intensities:



A crude local averaging technique is to set each value equal to the average of it and one value either side of it. Thus, the value 3 near the middle of the array is set to its average with its adjacent neighbours: ⅓ (4 + 3 + 6), which rounding to the nearest integer equals 4; and the 4 at the left-hand end of the strip is set to the mean of 4 + 5, which with rounding equals 4. This procedure is a form of weighting in which we start at one end of the array and work along it from left to right, adjusting each value according to a mathematical operation, yielding a local average. The application of a mathematical operation to an array is known as 'convolution'. Convolving the averaging operation with the array above yields to the nearest integer:

4 4 4 4 6 8 8 8 8

corresponding to the following graph:



Smoothing away the local irregularities begins to reveal that there may be a boundary between two different levels of intensity, one region at a level of 4 and one at a level of 8.

A more sensible averaging function takes account of a wider range of neighbours but weights them so that the further away they are, the smaller their contribution to the average. There are many possible operations, but a useful one is based on the *normal distribution*, which often arises in connection with measurements. Figure 4.4 shows its bell-shaped curve.

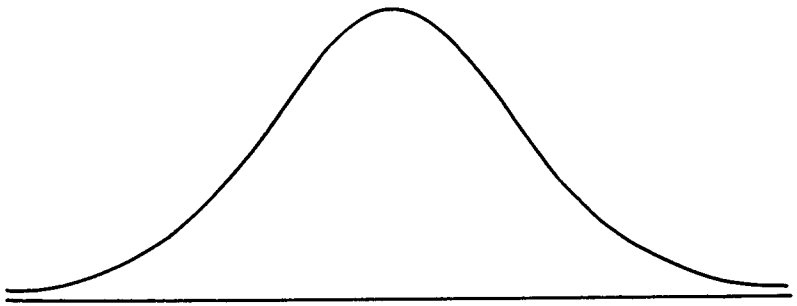
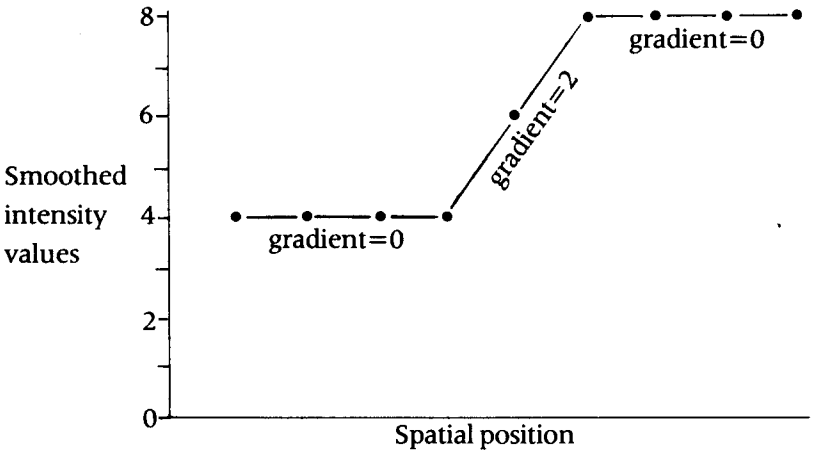
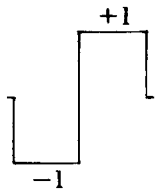


Figure 4.4: The bell-shaped curve of the normal distribution.

Given that the values in the grey-level array are smoothed out by local averaging, how are the intensity boundaries to be detected? The boundary between a bright region and a dull region corresponds to a relatively abrupt shift from values of one size to those of another size. A strip from the relevant part of the array will have values like those in the previous graph. If you were to walk along a surface supported by the bars in that graph, it would be flat at both ends but have a steep gradient in between:



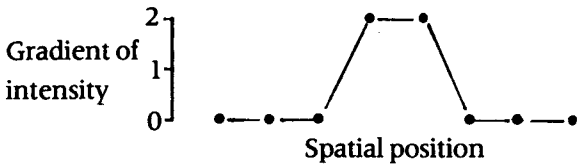
A simple way to measure the steepness of the gradient between any two adjacent values is to multiply the one on the left by −1 and the one on the right by +1, and to sum the results. Two adjacent intensity values near the middle of the graph are 4 and 6, and so the gradient between them is: $(4 \times -1) + (6 \times +1) = 2$. We can work along the array above applying this operation:



The resulting values of the gradient are as follows:

0 0 0 2 2 0 0 0

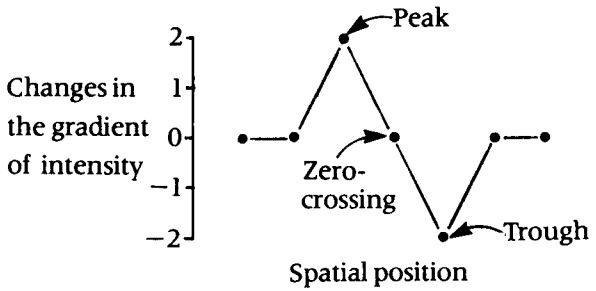
and we can plot them in a graph:



The boundary between the two regions of different intensity corresponds here to a local highpoint in the values of the gradient of intensity. The boundary also shows up in the *changes* in the gradient of intensity. At the left-hand end of the surface above, the gradient is constant, then its values go up (from 0 to 2), stay constant in the middle (from 2 to 2), and then go down (2 to 0). Finally, the slope is constant once again at the right-hand end of the array. Our operation above measures these changes in gradient, and it yields the values:

0 0 2 0 −2 0 0

which we can plot in a graph:



The point in the middle where the value is zero and the curve crosses from a peak of positive values to a trough of negative values is called a 'zero-crossing'. It and its adjacent peak and trough provide excellent evidence for the existence of a boundary between two regions of different intensities.

The operations of local averaging and finding changes in gradient have to be carried out in two dimensions, not just on a single one-dimensional strip. However, the two operations can be combined into one. This combined operation looks like a Mexican hat; Figure 4.5 shows a cross-section through it.

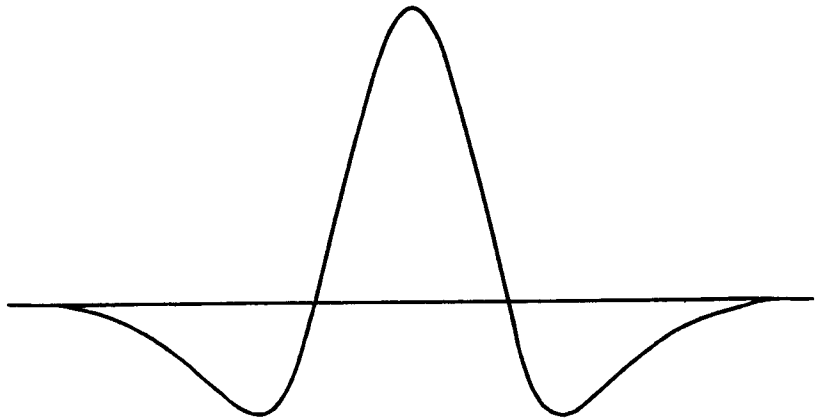
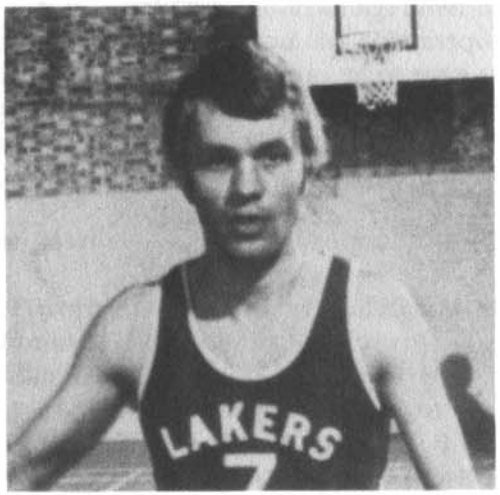


Figure 4.5: A cross-section through the 'Mexican hat' filter, which combines local averaging with the detection of changes in the intensity gradient.

Each value in the grey-level array is averaged with its neighbours weighted according to the Mexican hat. As the figure shows, the weights will be positive for near neighbours, decline to negative for more distant neighbours, and taper off to zero for points so distant that they need not be taken into account. When all the numbers in the grey-level array are filtered through the Mexican hat, the result is an array containing positive and negative values. The boundaries between these areas yield a map of the zero-crossings. Figure 4.6 shows a grey-level image and a map of the zero-crossings obtained from it.

Figure 4.6: (a) A grey-level image. (b) A map of zero-crossings: the intensity of the lines here varies with the contrast in intensity between the two regions on either side of the zero-crossing.



(a)

(b)

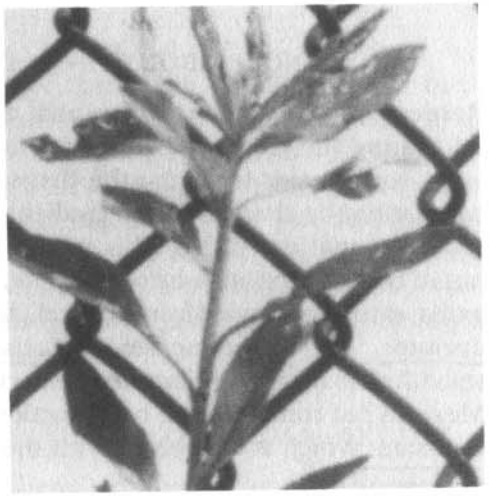
(From D. Marr, *Vision*. San Francisco: W. H. Freeman, 1982, p. 61)

VISUAL FILTERING

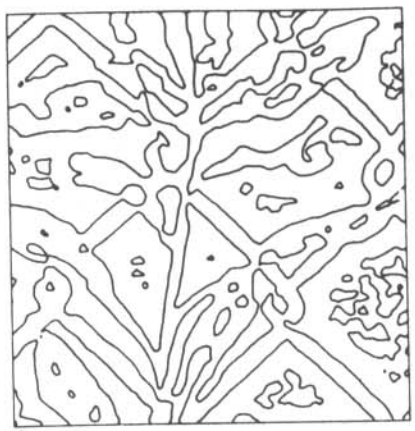
What size Mexican hat should be used in filtering the grey-level array? All sizes will detect sharp and clearly separated intensity changes. A large hat extending over many elements in the array will also reveal gradual changes in intensity over a large area,

Figure 4.7:
(a) A grey-level image (320 × 320 pixels).
(b) The zero-crossings obtained from it with a small filter (with an excitatory region of about 9 pixels in size).
(c) The zero-crossings obtained with a larger filter (with an excitatory region of about 18 pixels in size).

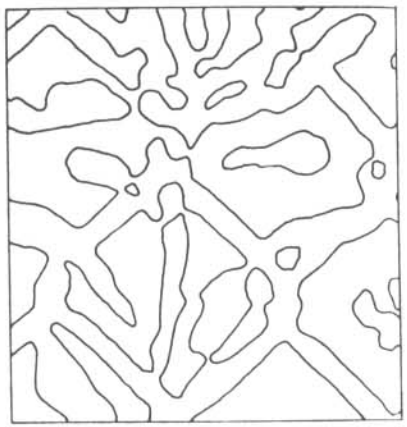
(From D. Marr, *Vision*. San Francisco: W.H. Freeman, 1982, pp. 58 and 72)



(a)



(b)



(c)

perhaps as a result of the illumination. A small hat extending over a few elements will reveal many small changes in detailed intensity. Since human beings are sensitive to a wide range of changes, we will design our robot to be similarly sensitive and to process the grey-level array with a series of different sizes of filter (i.e. Mexican hats). Figure 4.7 shows a picture (a grey-level image) and two maps of zero-crossings obtained from it, one using a small filter and the other using a filter of twice the size.

A NOTE ABOUT THE MATHEMATICS

Readers familiar with the differential calculus will recognize that calculating the gradient of intensity in the smoothed grey-level array corresponds to taking the first spatial derivative, and that calculating the changes in the gradient corresponds to taking the second spatial derivative. This second derivative can be carried out in two dimensions isotropically, i.e. giving equal weight to all paths extending away from a pixel, by the so-called Laplacian operator. The two-dimensional Gaussian normal distribution smooths the data, with importance decreasing with distance. The Mexican hat combines the two functions: it is this Laplacian of a Gaussian, which is convolved with the intensity array.

THE NEUROPHYSIOLOGY OF VISION

Trying to understand vision by studying only nerve cells, as Marr remarked, is like trying to understand bird flight by studying only feathers. One of the striking features of his account is that it fits together what the visual system does, how it does it, and which cells carry out the computations.

There are cells in the human retina (see Figure 4.8) which gather information from a set of retinal receptors lying in a roughly circular field. Some of these 'ganglion' cells are excited by light falling on the receptors in the centre of their field, and inhibited by light falling on those in the surround. Other ganglion cells have exactly the opposite characteristics: they are inhibited by light in the centre and excited by light in the surround. These two sorts of cells are just what are needed to carry out the

computations of the Mexican hat. The first sort calculate the positive values in the centre of the hat, and the second sort the negative values of the brim. Every nerve cell fires spontaneously from time to time. A cell normally registers significant information by departing from its spontaneous rate of firing, and cannot directly report both positive and negative values. They are therefore signalled by separate cells. Hence, zero-crossings correspond to locations where activity in the two sorts of ganglion cell is about equal.

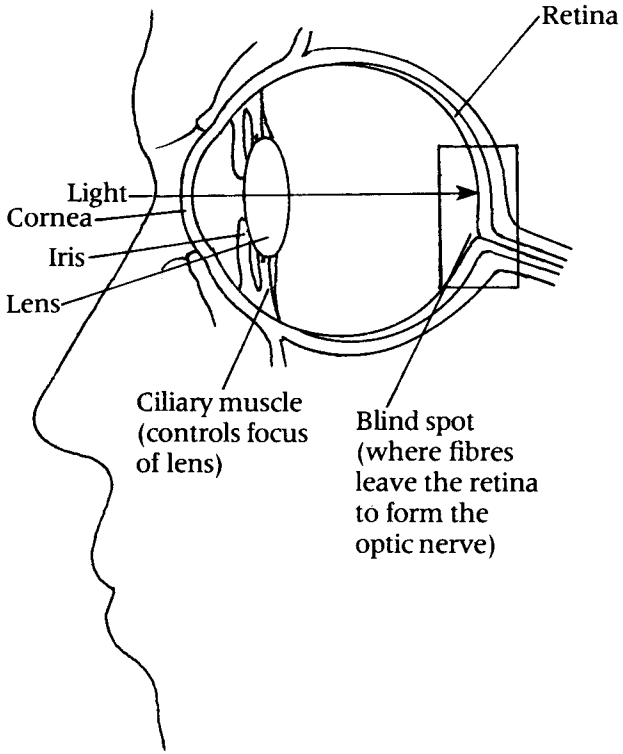
The neurophysiologists David Hubel and Torsten Wiesel, who received the Nobel Prize for their work, discovered cells in the visual cortex at the back of the brain that are excited by bright lines or bars at particular orientations in the visual field. One erstwhile interpretation of these cells is that they detect edges, borders and other features of objects. A more plausible interpretation is suggested by Marr's theory: they detect zero-crossings. They are excited by a line of neighbouring retinal ganglia, which in turn are excited by a change in the gradient of intensity aligned with their fields of receptors.

THE THIRD STAGE IN VISION: THE PRIMAL SKETCH

The grey-level array is filtered through a set of different sized Mexican hats. The results can be interpreted taking into account a fact that evolution is likely to have 'wired into' the brain: one thing cannot be in two places at the same time, whether it is an edge of an object, a change in the reflectance of a surface, or a change in illumination. Hence, if any one of these phenomena creates a zero-crossing in a filtered image, there is likely to be a corresponding zero-crossing not too far away in images produced by filters of other sizes. However, thin bars and other details may yield two zero-crossings from a small filter that become blurred together by larger filters. Similarly, two different phenomena may produce intensity changes at the same place but on different scales. A comparison of the filtered images is therefore highly informative. In general, where they concur, a zero-crossing is a result of the same physical phenomenon.

What the visual system extracts from a comparison of the

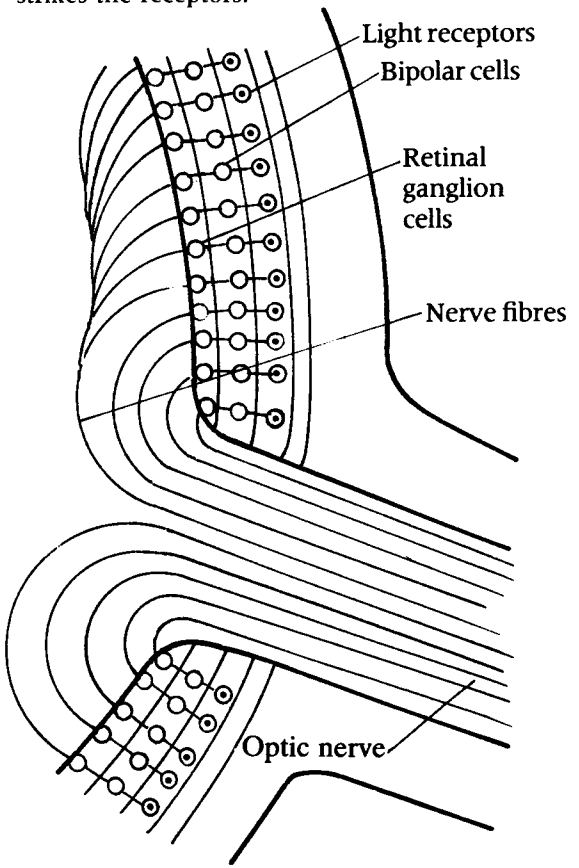
Figure 4.8(a): A diagram of the eye.



(A black and white version of Plate 9 from J. P. Frisby, *Seeing: Illusion, Brain and Mind.* Oxford: Oxford University Press, 1979)

filtered images is a matter of controversy. For Marr, as I have explained, the correspondences of zero-crossings are the critical data. But at curves and corners the zero-crossings from different sizes of filter do not lie in the same position in the array (see Figure 4.7). Moreover, as we shall see in the next chapter, human vision seems to be sensitive to the peaks and troughs adjacent to zero-crossings. Roger Watt and Michael Morgan have argued that some zero-crossings are the spurious results of noise,

Figure 4.8(b): A schematic detail from 4.8a showing the positions of the receptor cells and the retinal ganglion cells. The light passes through the nerves and ganglion cells before it strikes the receptors.



(A black and white version of Plate 9 from J. P. Frisby, *Seeing: Illusion, Brain and Mind.* Oxford: Oxford University Press, 1979)

and that it is better to rely on peaks and troughs. A program that they have devised keeps the positive and negative values of the filtering process separate, working out their averages from different sizes of filter. It locates the centres (and sizes) of the peaks in the positive averages and the centres (and sizes) of the troughs in the negative averages, which it then uses to construct

a symbolic representation of bars, edges and regions of the same intensity.

Bars, edges and blobs are the basic elements out of which the visual image is constructed. They are the sorts of primitive that artists who etch pictures have at their disposal. Each has a specified position, orientation, length, width and contrast in intensity with its surrounding region. Marr makes this information explicit by using symbolic descriptions with appropriate numerical values, such as:

BLOB
(POSITION 146 21)
(ORIENTATION 105)
(CONTRAST 76)
(LENGTH 16)
(WIDTH 6)

They are attached to locations in the map summarizing the results of filtering the grey-level array. This particular description applies to the blob below the wire in the top left-hand corner of Figure 4.7(b).

Such information captures the local details in the visual image, and provides the raw data for what Marr called the 'primal sketch', which makes explicit the complete organization of the visual image – roughly, what you are aware of when you screw up your eyes and look at the world slightly out of focus. The primal sketch is constructed by grouping similar elements together to form lines, larger blobs and structured groups, and the process is progressively repeated on an ever larger scale. These principles, however, have yet to be implemented in a program. They are difficult to isolate because they organize visual images automatically with no conscious effort. Thus, Figure 4.9 seethes with potential groupings, and as soon as a potential organization arises it is succeeded by a rival candidate, particularly if you allow your eyes to wander from place to place. The primal sketch should reveal what is crucial to locating discontinuities in the physical surfaces of the things in the scene. Once again, the underlying principles are not well understood, but presumably depend on inbuilt assumptions about the world.

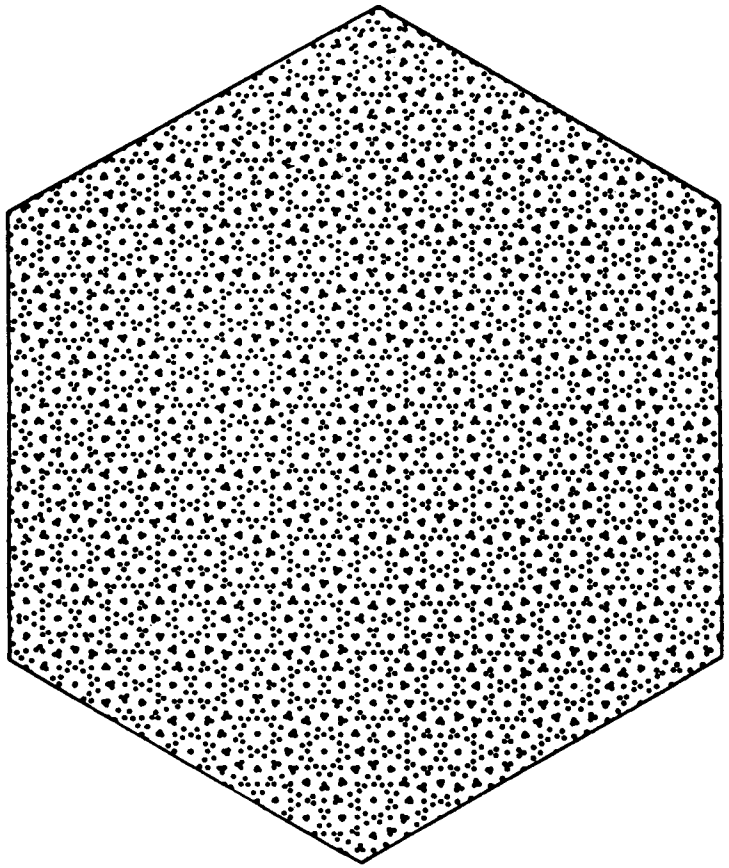


Figure 4.9: A figure with many potential organizations.

(From J. L. Marroquin, Human visual perception of structure. Master's thesis, Department of Electrical Engineering and Computer Science, MIT, 1976)

When you view things through half-closed eyes, you lose information. Paradoxically, this manoeuvre can help you to see something otherwise not apparent. The checkerboard picture in Figure 4.10, which was devised by Leon D. Harmon, reveals an identifiable person only if you blur its image by squinting at it from a distance. This procedure reduces the information about the sharp edges of the blocks, which otherwise interferes with the interpretation of the large-scale changes in intensity. They suffice for you to recognize Abraham Lincoln.

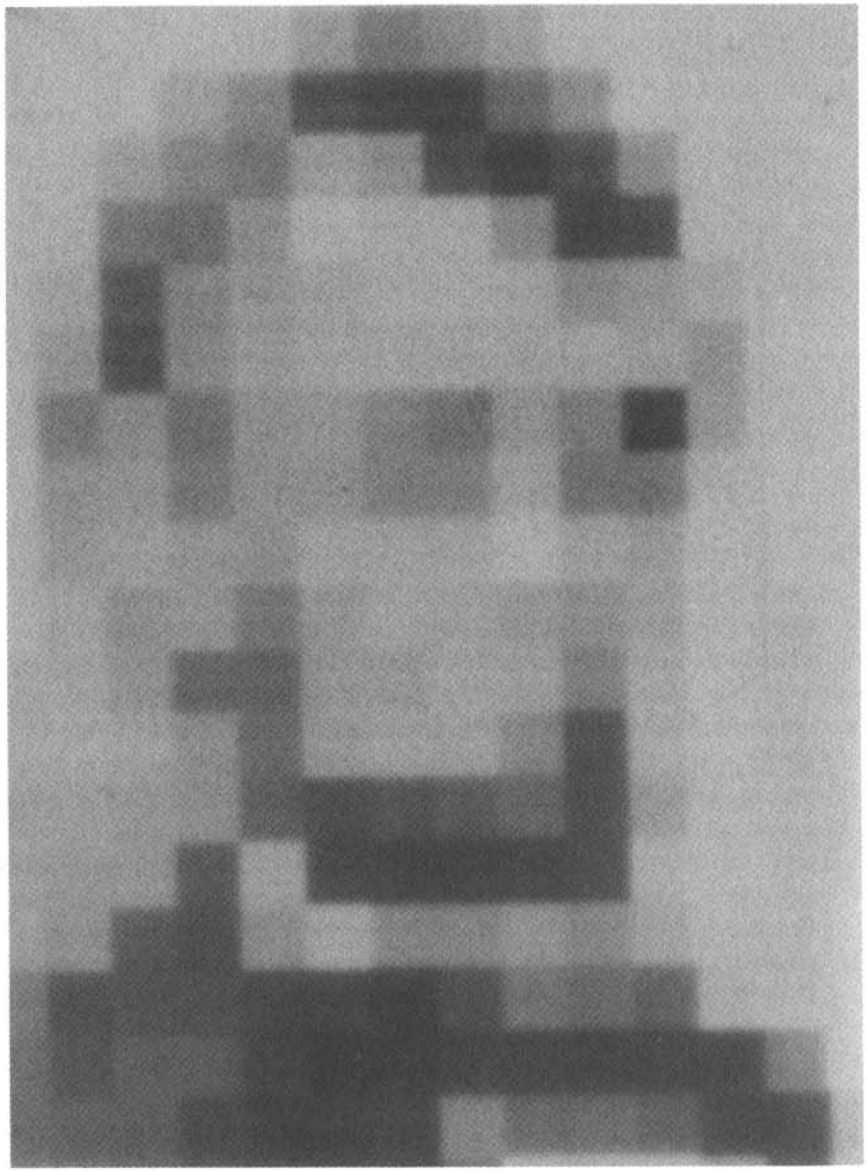


Figure 4.10: A checkerboard picture devised by Leon D. Harmon. It should be viewed at a distance through half-closed eyes.

(From L. D. Harmon, The recognition of faces. *Scientific American*, November 1973, p. 75)

THE COST OF EARLY VISUAL PROCESSING

Present-day technology can equip a robot with an eye containing about a million light-sensitive elements, though they take up a larger area than the much more densely packed human retina. Nature has mastered small-scale wiring in three dimensions and the parallel execution of myriad computations. Each cell in the nervous system has thousands of connections to other cells and carries out its own computations. Electronic components such as the micro-chip, however, are restricted to an essentially two-dimensional wiring of components, with far fewer interconnections than occur between nerve cells. Most digital computers carry out just a single sequence of computations, and parallel computers are a recent innovation. Hence, the major problem in equipping a robot with a program to construct the primal sketch is to ensure that it works fast enough. A computation that takes two seconds is no use if you are travelling at fifteen miles an hour. By the time you are aware of a large black patch in the image, you will be falling down the hole that caused it.

The major computational cost is the filtering of the grey-level array. Each pixel in the 1000 × 1000 array produced by an electronic camera has to have a mathematical operation applied to it that takes into account the values of its neighbours, and this process has to be repeated for all the different sizes of filter. Special electronic devices have been built to carry out this operation, but have yet to rival the efficiency of the nervous system.

A robot whose vision of the world was solely a primal sketch would be comparable to a common housefly. Despite its amazing aerial agility, the housefly probably does not form a three-dimensional model of the world. Its flight pattern, as Werner Reichardt and his colleagues at Tubingen have shown, is controlled by rapid and automatic mechanisms. One mechanism puts the fly into its landing routine if its visual field suddenly expands at a high speed. Wherever the looming surface is, the fly turns its feet towards the centre of expansion, and as soon as they touch the surface, the power to its wings is cut off. Another

mechanism allows the fly to track its mate. This system is sensitive to a small black patch moving against a background. The relative power provided to the left and right wings is governed by the patch's position in the visual field and by its angular velocity, and the fly thereby keeps the target centred in its visual field and flies towards it. The real size of the target, however, is not critical. All that matters is its size in the visual image. Thus, the argument that vision is impossible, which I outlined at the start of the chapter, applies to the fly. It cannot tell the difference between a nearby mate and a distant passing bird if they each project an image of the same size on its eye. However, the mechanism controlling its flight path is innately tuned for small, nearby objects. If the fly attempts to intercept a distant bird, it will fail; the converse, alas, may not be true.

The fly's visual system is adapted to its world. It needs to home in on mates, to land on surfaces, and to take off rapidly if another looming surface hurtles towards it. If our robot is to inhabit a richer world, it will have to perceive it in three dimensions just as people do. It will have to treat the visual image – the information made explicit in the primal sketch – as a precursor to the recovery of information about the physical surfaces giving rise to it.

Further reading

Poggio (1984) is an excellent short introduction to Marr's work. Frisby (1979) has written an attractive book on vision, full of splendid illusions and embued with the computational approach. Mayhew and Frisby (1984) give a more technical account of computer vision. Watt (1988) is an advanced monograph on the initial stages of human vision.

from a particular point of view. If you look at your hand edge-on with your fingers closed, then several fingers may yield a coinciding silhouette at various parts. Likewise, if you turn your hand away slightly, then adjacent points of its silhouette may belong to two quite different fingers, one behind the other. And, of course, its silhouette will not consist of points that all lie in the same plane. In general, when such violations occur, it is impossible to see the true three-dimensional shape of an object from its silhouette unless it *is* as familiar as the back of your hand. Where the assumptions do hold, however, then the visual system can usually perceive the true shape of an object.

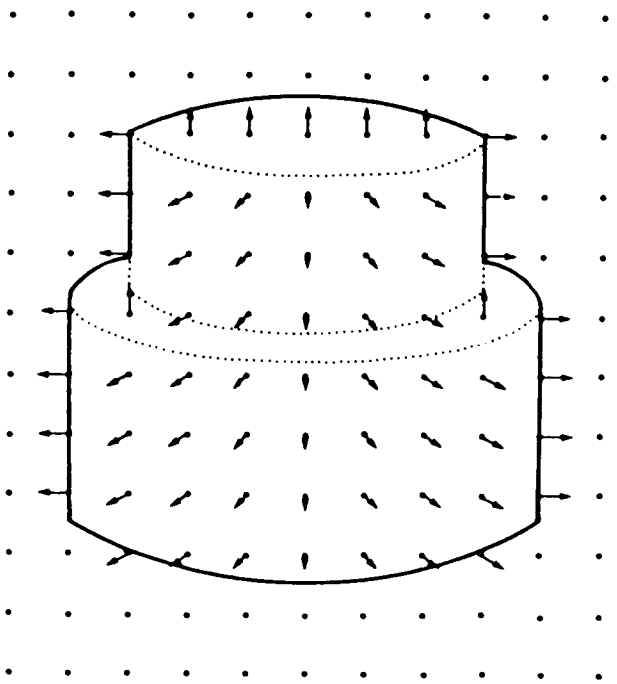
THE TWO-AND-A-HALF-DIMENSIONAL SKETCH

Whenever the mind accomplishes a seemingly impossible task, it must have independent information about the world. Vision appears to be impossible, but evolution has equipped the visual system with inborn constraints that enable it to cope. Uniqueness, continuity of surface, rigidity and the constraints on contours all play a role in recovering the depth and orientation of surfaces from the primal sketch. There are other cues that are biologically important, e.g. the brightness and hue of a surface, and its shading and surface contours. They too probably rely on such built-in constraints, which operate automatically and unconsciously, and are relatively unaffected by a knowledge of objects. They are part of what one might call *pure* perception: the set of visual modules that operate in bottom-up mode to transform the grey-level array into the primal sketch, and that then carry out stereopsis, interpret motion, recover shape from contour, and perform all the other processes underlying the perception of surfaces.

The cornerstone of Marr's theory of vision is that the last stage of pure perception delivers an explicit representation of the relative depths and orientations (with respect to the observer) of each visible surface in a scene. He calls this representation the 'two-and-a-half-dimensional sketch' (or 2½ D sketch). The expression should not be taken literally: it merely designates a representation that does not make explicit the full three-dimensional

relations among objects, since the depths are relative to the observer. The representation is sometimes depicted in a diagram that looks like a pin-cushion in which each pin represents the depth and orientation of a region of a surface (see Figure 5.15). The sources of the two-and-a-half-dimensional sketch are stereopsis, motion, contour and all the other depth cues that I have discussed in this chapter. It is supposed to integrate the information that they yield, establishing consistency and filling in missing parts of surfaces. Whether, in fact, the human visual system constructs such a representation is not known.

Figure 5.15: An illustration of the two-and-a-half-dimensional sketch.



(From D. Marr, *Vision*. San Francisco: W. H. Freeman, 1982, p. 129)

Although computer programs for many of these sources of information have been implemented, it is not yet possible – unfortunately for our robot – to model all the processes that

enable us to see the world in depth. Moreover, the two-and-a-half-dimensional sketch is not sufficient for a robot to navigate safely through the world, since it makes explicit only the appearance of surfaces from the robot's point of view. A sensible representation of a scene, as we shall see in the next chapter, should enable objects to be identified, and therefore needs to be independent of any particular point of view. Identification is no longer a matter of pure perception, since it depends on assumptions that are the result of a person's own experiences.

Further reading

Marr (1982) and Ullman (1979) are the best sources for the topics of this chapter. Special issues of the journals *Artificial Intelligence* (1981, Vol. 17) and *Cognition* (1984, Vol. 18) have been devoted to vision. Koenderink (1984) gives an advanced account of the role of contour as a cue to shape, and also points out errors in Marr's analysis. Brady (1983) considers the analysis of shape from the standpoint of machine vision. Marr's notion of a visual module, as exemplified by the mechanism for stereopsis, has led Fodor (1983), along with other considerations, to propose a modular architecture for much of the mind.

CHAPTER 6

Scenes, shapes and images

Our goal is to understand human vision and to implement, if possible, a similar system in a robot. What needs to be computed is a symbolic representation of the three-dimensional world that makes explicit *what* is *where*. You walk into a room, recognize that it contains a table and other items of furniture, and can make your way to the table. You can carry out this action even if you have never been in the room before. There are three problems that your visual system solves: it perceives the three-dimensional shapes of objects, it identifies the objects on the basis of their shapes (the *what* in the symbolic representation), and it perceives their relative locations in space (the *where* in the symbolic representation).

The perception of the shapes of objects and of the spatial relations among them are not two independent tasks, but the same task on two different scales: a scene is simply a complex object made up of many component objects, and most objects in turn are made up of component parts, and so on and on. Just as one object can move in relation to another, so one part of an object can move in relation to another, as when you open and close your hand. Where a difference emerges is that objects (and their parts) tend to have names and functions, but few scenes (with the exception of rooms) have either. I shall begin with the perception of shapes, and defer the naming of parts until later.

THE CONSTRUCTION OF A THREE-DIMENSIONAL MODEL OF THE WORLD

Suppose that a robot has computed a symbolic representation of the relative depths and orientations of all surfaces lying within its view – the two-and-a-half-dimensional sketch that I described in the previous chapter (though some modern robots cheat by using

lasers to measure these distances). This representation does not make explicit the shapes of objects or their spatial relations, and it changes when the robot moves, because the orientations of the surfaces change in relation to the robot. A more useful and stable representation would make explicit the intrinsic three-dimensional shapes of objects and the spatial relations among them.

Confronted with the scene in Figure 6.1 – a seventeenth-century world of blocks similar to a domain beloved by workers in artificial intelligence – the robot needs to construct a representation that is independent of its particular point of view and that would enable the program controlling its movements to ensure that it does not collide with any obstacles. The representation must therefore be a *model* in three dimensions of the scene, which, like an architect's model, makes explicit the shape of everything in the scene – the regions that are filled and the empty spaces. Such a structure does not call for a three-dimensional layout in the hardware of the computer. Its physical embodiment merely has to function as three-dimensional so that elements can be accessed and manipulated by specifying their positions on three coordinates. Such structures are commonplace in programs.

The construction of the model depends on a geometrical transformation of the two-and-a-half-dimensional sketch. A similar task arises in constructing computer models of terrains from radar scans or from stereo photographs. There are programs that convert such data into three-dimensional models within the computer, and that can present projections of the scene from different viewpoints. The programs are relatively straightforward, but how the human visual system carries out the analogous transformations is not known.

THE INTERPRETATION OF LINE DRAWINGS

A number of workers in artificial intelligence, notably the late Max Clowes, David Huffman and David Waltz, set out to solve an analogous problem. They noted that line drawings make explicit the edges of objects, and that people interpret such drawings as

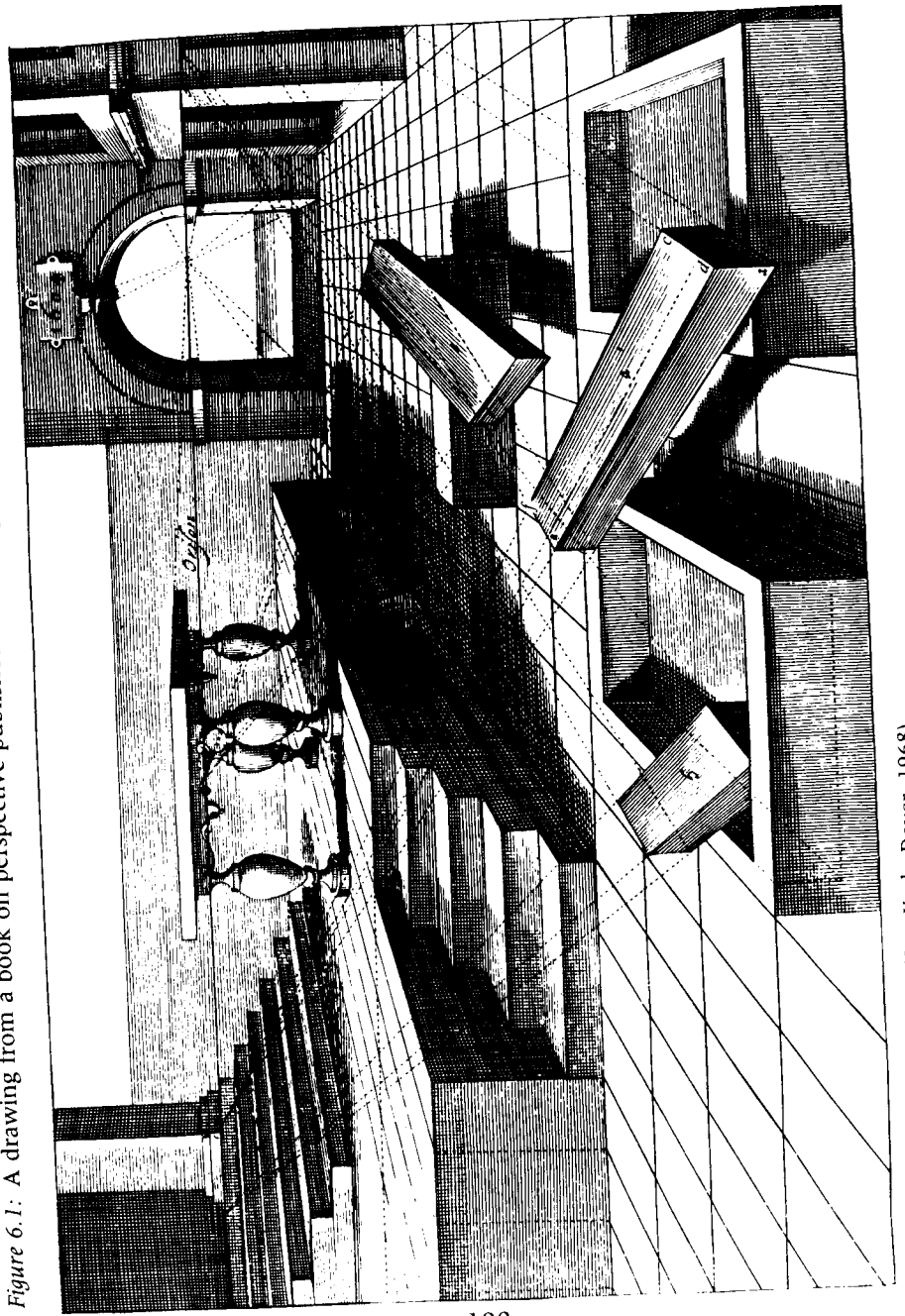


Figure 6.1: A drawing from a book on perspective published in 1604 by the Dutch artist Jan Vredeman de Vries. (From J. V. de Vries, *Perspective*. New York: Dover, 1968)

depicting three-dimensional scenes. They attempted to devise programs that could make similar interpretations.

What Clowes and Huffman independently realized is that knowledge of the world constrains the interpretation of the primitive symbols in a line drawing, and so enables a sensible three-dimensional interpretation to be made. Since their ideas are similar, I shall describe only Clowes's program. It takes as its input line drawings of a simple world of blocks. Each block has plane-faced surfaces and only three surfaces can meet at a corner. The program delivers an interpretation of the drawing in which each primitive symbol (lines and junctions of lines) receives a label representing its appropriate three-dimensional interpretation. At the heart of the program is a dictionary that defines the set of possible meanings for each sort of primitive symbol that can occur in a drawing. A single straight line, such as:

has only four possible meanings. It may stand for an external edge that bounds an object in the scene. If so, one side of the line corresponds to a surface of the object and the other side corresponds to the background or to some surface occluded by the object. There are obviously two interpretations here, depending on which side of the line the surface of the object lies. But the line may also depict an edge that is the meeting place of two surfaces of a single object. In this case, the edge can be either convex as on the top of a block or concave as on the inside of a box.

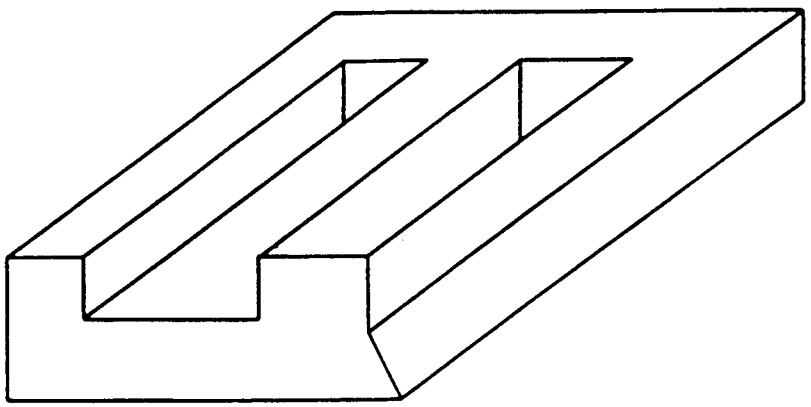
In drawings of the Clowes–Huffman world of blocks, only four sorts of junctions between lines can occur: an L-shape, a T-shape, a Y-shape and an arrow-shape. They can be seen on the beam lying across the boxes in Figure 6.1: there is an L-shape at the point labelled *f*, T-shapes at each intersection between the line depicting the bottom of the beam and the lines depicting the top of a box, a Y-shape at the point labelled *d*, and an arrow-shape at the point labelled *c*.

Since there are four possible interpretations for a single line, you might imagine that there should be sixteen possible interpretations for an L-junction (the four for one line multiplied by the

four for the other). In fact, many of the combinations are ruled out as nonsensical by the nature of the blocks world. For instance, at least one of the two lines in an L-junction must denote the edge of an object that occludes another surface or the background. There is a similarly constrained set of interpretations for the other sorts of junctions.

The interpretation of a drawing exploits a higher level form of constraint: the need for a consistent interpretation of all the primitive symbols (lines and junctions) in the drawing. Even if each primitive in a drawing is initially assigned only physically possible interpretations, there will still be a massive number of potential interpretations of the drawing as a whole. Most of them, however, will denote objects that are impossible, such as the one in Figure 6.2. If the interpretation of a junction treats one line as an occluding edge of an object, then that line must also be an occluding edge in the interpretation of the junction at its other end. The object in Figure 6.2 violates this constraint: if you look at the front of the object alone, or its back alone, what you see is sensible. The two parts, however, call for incompatible interpretations of the lines connecting them.

Figure 6.2: A line drawing of a figure that is recognized as physically impossible by Max Clowes's program for interpreting drawings.

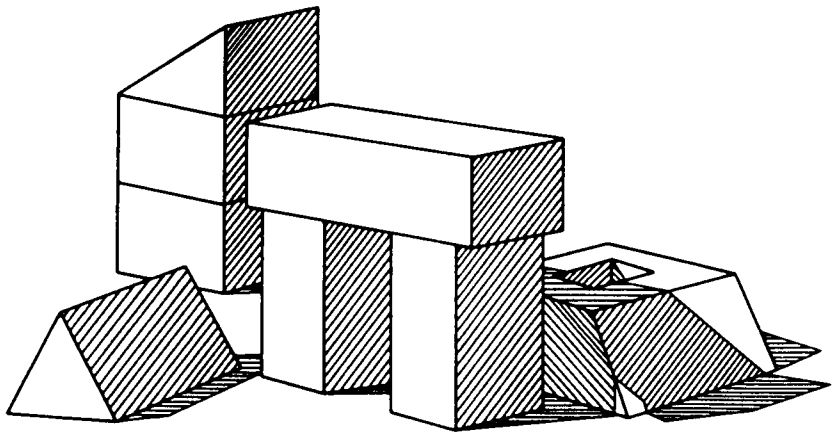


(From M. Clowes, On seeing things. *Artificial Intelligence*, 2, 1971, p. 105)

The procedure for interpreting the drawings first assigns all legal interpretations to each primitive, and then checks the consistency of the interpretations of a neighbouring pair. When it has eliminated all but the consistent ones, it considers the interpretations of another neighbouring primitive, establishes a consistent interpretation for all three, and so on, until every primitive in the drawing has been pulled into the process. When the program finally stabilizes – it is using the 'relaxation' method discussed in Chapter 5 – the result will be those interpretations that make sense of the drawing. If, in fact, it denotes an impossible object then there will be no resulting interpretation for it.

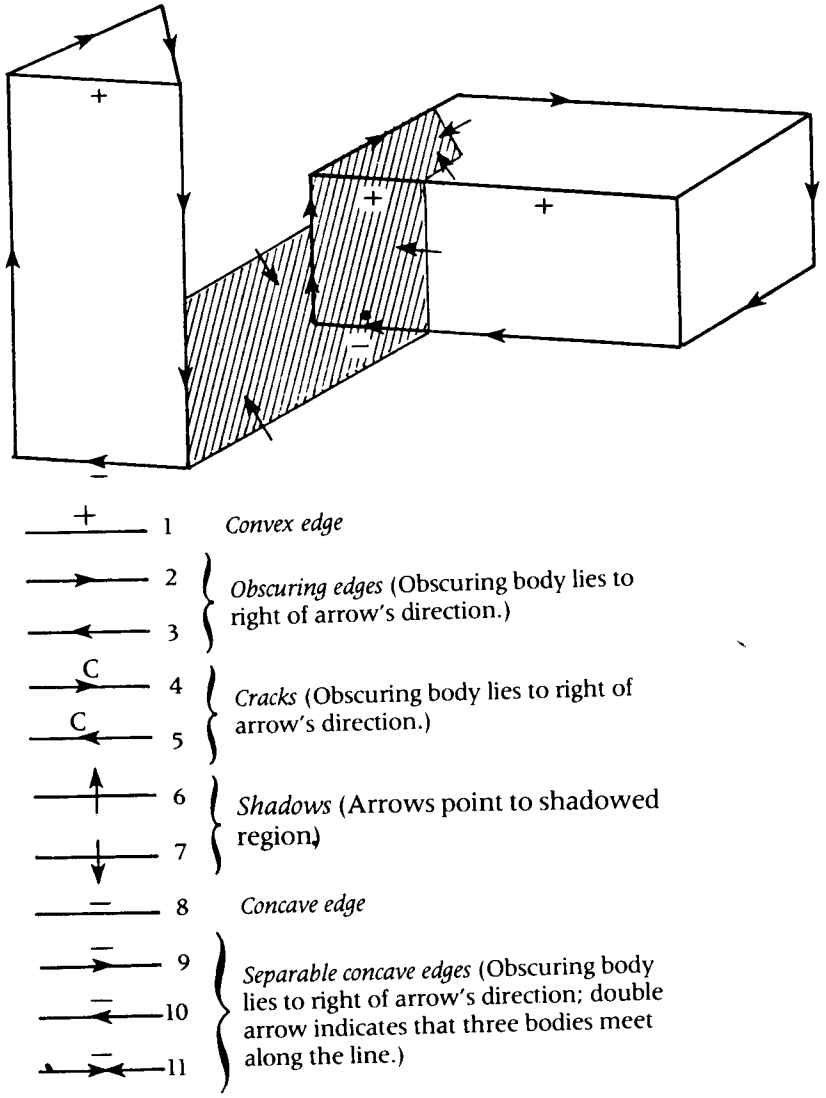
David Waltz showed that a step that seemed to complicate matters – the introduction of shadows, more complex blocks, and more complex configurations of blocks (see Figure 6.3) – led in fact to a simplification. His program allows that a line may denote a crack where one block lies next to another or on top of a surface, or that it may denote the edge of a shadow with shade on one side and light on the other. Shadows give information that is

Figure 6.3: A line drawing of blocks to be interpreted by a computer program.



(From D. Waltz, Understanding line drawings of scenes with shadows. In P. Winston, ed., *The Psychology of Computer Vision*. New York: McGraw-Hill, 1975)

Figure 6.4: The labelling of a line drawing according to David Waltz's program.



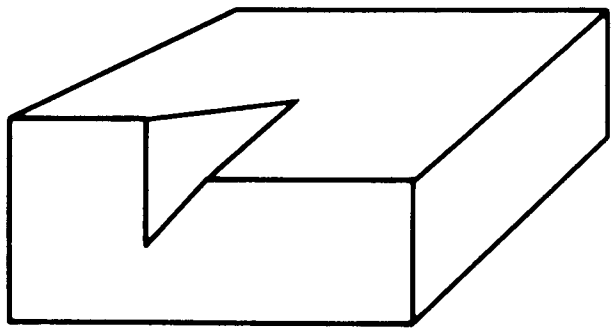
(From D. Waltz, Understanding line drawings of scenes with shadows. In P. Winston, ed., *The Psychology of Computer Vision*. New York: McGraw-Hill, 1975)

analogous to another viewpoint, and can reveal whether an object is resting on a surface or is merely close to it. Figure 6.4 illustrates some of these possibilities in a labelling of a drawing according to Waltz's program. Although there are now more primitive symbols and many more meanings for them, the program arrives at an interpretation of a drawing as a whole more rapidly (and less ambiguously) than in the simpler blocks world.

Other workers, such as Alan Mackworth, have devised still more advanced programs, which can label drawings using general principles rather than a dictionary of interpretations for individual primitives. Yet there remain problems in working within a blocks world. The programs give a sensible interpretation to certain drawings that depict impossible objects (see Figure 6.5). Marr criticized the approach on the grounds that it fails to deal properly with the question of what should be computed. The output of a program, if it is to resemble human vision, should be a three-dimensional interpretation of the scene depicted in the drawing. Such a representation needs to make explicit the shapes of objects, but the programs that label lines recover only the orientations of connected surfaces.

Nevertheless, the research was valuable. It clarified the distinction between form and function. Thus, Stuart Sutherland,

Figure 6.5: A line drawing of a figure that is not recognized as physically impossible by programs for interpreting line drawings.



(From J. Mayhew and J. Frisby, Computer vision. In T. O'Shea and M. Eisenstadt, eds., *Artificial Intelligence*. London: Harper and Row, 1984)

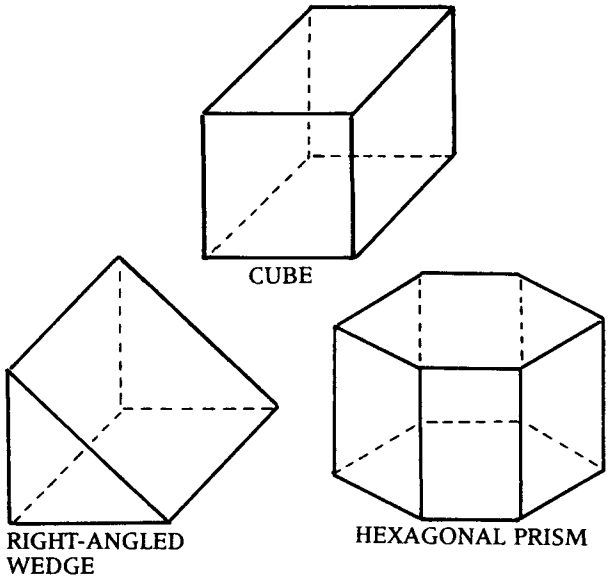
who had studied the way in which animals discriminate shapes, distinguished three separate domains: the domain of the picture, e.g. lines, regions and junctions; the domain of the scene, e.g. surfaces, edges and shapes; and the domain of functional objects, e.g. chairs, tables and people. The research also convinced many psychologists that there were mundane aspects of perception that could be investigated without having to carry out elaborate experiments: simple intuitions about drawings were enough to raise questions that no one knew how to answer.

THE IDENTIFICATION OF OBJECTS

The construction of a mental model can call for more than a transformation of the two-and-a-half-dimensional sketch. There are often insufficient data in the sketch for a complete model. For instance, only two edges of the table-top in Figure 6.1 are visible, yet if you are familiar with Western furniture you will identify the object as a table and represent its top as rectangular. You depend on your knowledge of the shapes of tables gained from your experience of the world, but the machinery of identification is unconscious in the Helmholtzian sense. Your visual system constructs a description of the perceived object and compares it with some sort of mental catalogue of the three-dimensional shapes of objects. It can recognize them from particular viewpoints and then make automatic extrapolations about the rest of their shapes.

The pioneer of computational models of the process was L. G. Roberts, who devised a program that interprets photographs of objects in a blocks world by identifying them with stored prototypes. Artists such as Cézanne and the Cubists were inspired by the Platonic doctrine that all shapes could be decomposed into a primitive vocabulary of elementary solid forms; and Roberts's program works on a similar principle. It is based on three space-filling prototypes: the cube, the right-angled wedge and the hexagonal prism (see Figure 6.6). It first converts a photograph into a line drawing by using a crude line-finder – a rudimentary precursor to the machinery of the primal sketch. It then works bottom up using particular junctions of lines as cues to a relevant

Figure 6.6: The three prototypes used in L. G. Roberts's program.



(From K. Oatley, *Perceptions and Representations*. London: Methuen, 1978, p. 178)

prototype. Thus, a Y-junction cues the prototype of the cube, since it occurs at a top corner on the front of the cube.

The final stage of the program uses the prototype in a top-down process to interpret the rest of the object in the drawing. The operations that carry out this matching process can project the internal prototype (specified in terms of coordinates in three dimensions) on to two-dimensional images, and they can stretch or shrink the dimensions of the prototype, rotate it, or translate it from one position to another, so as to fit the actual object in the scene. The program can also join prototypes together in order to analyse more complex objects.

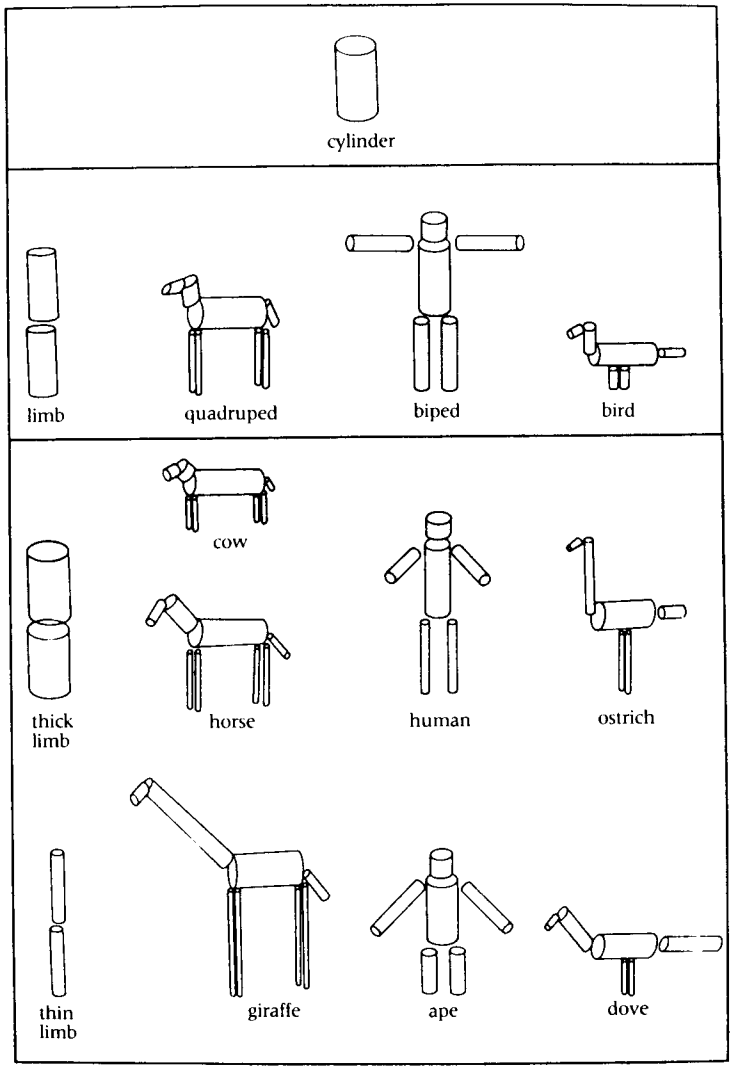
The use of prototypes was taken a step further by David Marr and Keith Nishihara. Since the identification of an object can occur from many different viewpoints, they argued that the shape of the object must be specified, not in a coordinate system centred on the observer (such as the two-and-a-half-dimensional sketch), but in coordinates that are determined by the shape of

the object itself. The visual system creates a representation of a folded umbrella that makes explicit that it is a long cylinder. This model is independent of viewpoint, and indeed people can imagine the appearance of an umbrella from different points of view. This ability to rotate mental images is one that I shall come back to presently.

A powerful system for capturing the shapes of objects is based on the idea of moving a two-dimensional cross-section along an axis. If a circle, for example, is moved at right-angles along a straight line, it sweeps out the solid shape of a cylinder. If it shrinks continuously as it moves along, it defines a cone. In general, we can use a cross-section of any shape, it can change size continuously as it moves, and the axis need not be a straight line but can curve in any way. The resulting class of shapes are known, somewhat misleadingly, as 'generalized cones'. Thus, the shape of a banana can be described reasonably accurately by a generalized cone, and the shape of a human being can be described reasonably accurately as made up from a number of generalized cones. Marr and Nishihara proposed that the mind contains a catalogue of shapes of objects represented as generalized cones. Figure 6.7 shows the sort of catalogue they envisaged, which works on the same representational idea as the 'stick figures' of children's drawings. It makes explicit the lengths and arrangements of the component axes, which is why it is useful for the identification of complex objects. Although Figure 6.7 represents the shapes as simple cylinders, the mental catalogue is supposed to employ generalized cones. They are the class of shapes that are recovered from silhouettes meeting the constraints described in the previous chapter. However, there are certain shapes, ranging from Origami to crumpled newspapers, that violate those constraints and that are not generalized cones.

Marr and Nishihara's catalogue represents complex objects in a hierarchical organization. The top level captures their gross structure of objects; the lower levels make the details explicit. Figure 6.8 shows the hierarchical representation of the shape of a human being. (Cylinders are again used for illustrative convenience.) The axis of the body provides a coordinate system centred on the body itself, and this system can be used for

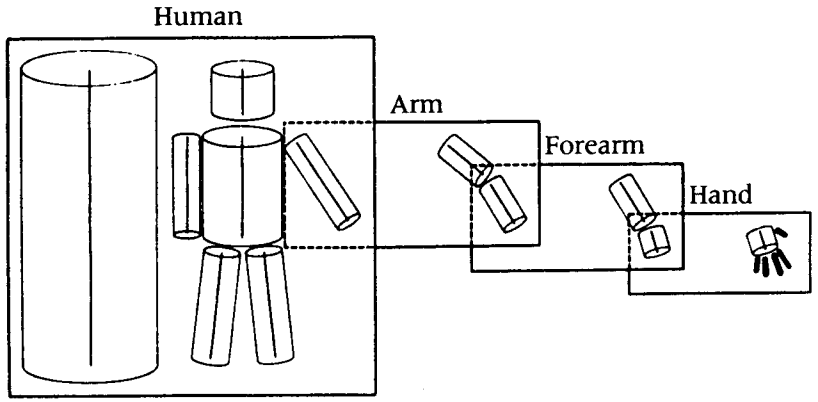
Figure 6.7: A fragment of the catalogue of shapes proposed by Marr and Nishihara.



(From D. Marr and H. K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society*, B, 1978, 200)

specifying the relations of the axes of the head, arms and legs. Their axes in turn provide coordinates for the next level down, and so on. The catalogue can be searched in a variety of ways. Like Roberts's program, a specific cue about a key part of an object may provide bottom-up access to a prototype. The prototype can then be used top-down to try to match the rest of the figure. Specific information about the orientation of the principal axis in relation to other axes may also be used in the matching process.

Figure 6.8: The hierarchical organization of a human being's shape according to Marr and Nishihara.



(From D. Marr and H. K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society*, B, 1978, 200)

David Hogg has extended some of these ideas to cope with movement. He has developed a program that interprets films of a man walking. The program has an internal prototype of a man, like the one in Figure 6.8, which it can project on to the moving image using principles like those of Roberts's program. The prototype also contains constraints on the variables that control the various angles of the joints so that the man can only cycle through the specified sequence of postures that occur in walking. The program operates on the raw data in the grey-level array,

and the matching process is triggered by the detection of a difference between successive frames of the film. In essence, the program draws a rectangle around the area of change and assumes that its major axis corresponds to the major axis of the prototype. Once it has established this relation, it switches to a top-down mode of operation in order to fit the details of arms and legs to the picture, allowing for the camera's point of view.

The theme of these approaches to recognition is the use of high-level knowledge of the shapes of objects. The theories assume that your experience enables you to build up such knowledge and to exploit it in procedures that make sense of the visual world. Is the assumption warranted? The reader will recall that earlier I described a relevant experimental tactic: if a process

Figure 6.9: A picture illustrating the role of top-down processes in perception.

(From J. Thurstone and R. G. Carraher, *Optical Illusions and the Visual Arts*. New York: Litton, 1966. Copyright © Van Nostrand Reinhold Co.)

occurs even though the low-level data are corrupted, then it can be guided by higher-level knowledge. Figure 6.9 is a classic illustration of just such a phenomenon. At first glance, the figure may appear to be random patches of black and white. It is, in fact, a picture of a dog sniffing the ground under the dappled shadow of a tree. This knowledge should be enough for you to identify the dog.