Vision by Man and Machine

5

How does an animal see? How might a computer do it? A study of stereo vision guides research on both these questions. Brain science suggests computer programs; the computer suggests what to look for in the brain.

• • •

Tomaso Poggio
April, 1984

The development of computers of increasing power and sophistication often stimulates comparisons between them and the human brain, and these comparisons are becoming more earnest as computers are applied more and more to tasks formerly associated with essentially human activities and capabilities. Indeed, it is widely expected that a coming generation of computers and robots will have sensory, motor and even "intellectual" skills closely resembling our own. How might such machines be designed? Can our rapidly growing knowledge of the human brain be a guide? And at the same time can our advances in "artificial intelligence" help us to understand the brain?

At the level of their hardware (the brain's or a computer's) the differences are great. The neurons, or nerve cells, in a brain are small, delicate structures bound by a complex membrane and closely packed in a medium of supporting cells that control a complex and probably quite variable chemical environment. They are very unlike the wires and etched crystals of semiconducting materials on which computers are based. In the organization of the hardware the differences also are great. The connections between neurons are very numerous (any one neuron may receive many thousands of inputs) and are distributed in three dimensions. In a computer the wires linking circuit components are limited by present-day solid-state technology to a relatively small number arranged more or less two-dimensionally.

In the transmission of signals the differences again are great. The binary (on-off) electric pulses of the computer are mirrored to some extent in the all-or-nothing signal conducted along nerve fibers, but in addition the brain employs graded electrical signals, chemical messenger substances and the transport of ions. In temporal organization the differences are immense. Computers process information serially (one step at a time) but at a very fast rate. The time course of their operation is governed by a computer-wide clock. What is known of the brain suggests that it functions much slower but that it analyzes information along millions of channels concurrently without need of clock-driven operation.

How, then, are brains and computers alike? Clearly there must be a level at which any two mechanisms can be compared. One can compare the tasks they do. "To bring the good news from Ghent to Aix" is a description of a task that can be done by satellite, telegraph, horseback messenger or pigeon post equally well (unless other constraints such as time are specified). If, therefore, we assert that brains and computers function as information-processing systems, we can develop descriptions of the tasks they perform that will be equally applicable to either. We shall have a common language in which to discuss them: the language of information processing. Note that in this language descriptions of tasks are decoupled from descriptions of the hardware that perform them. This separability is at the foundation of the science of artificial intelligence. Its goals are to make computers more useful by endowing them with "intelligent" capabilities, and beyond that to understand the principles that make intelligence possible.

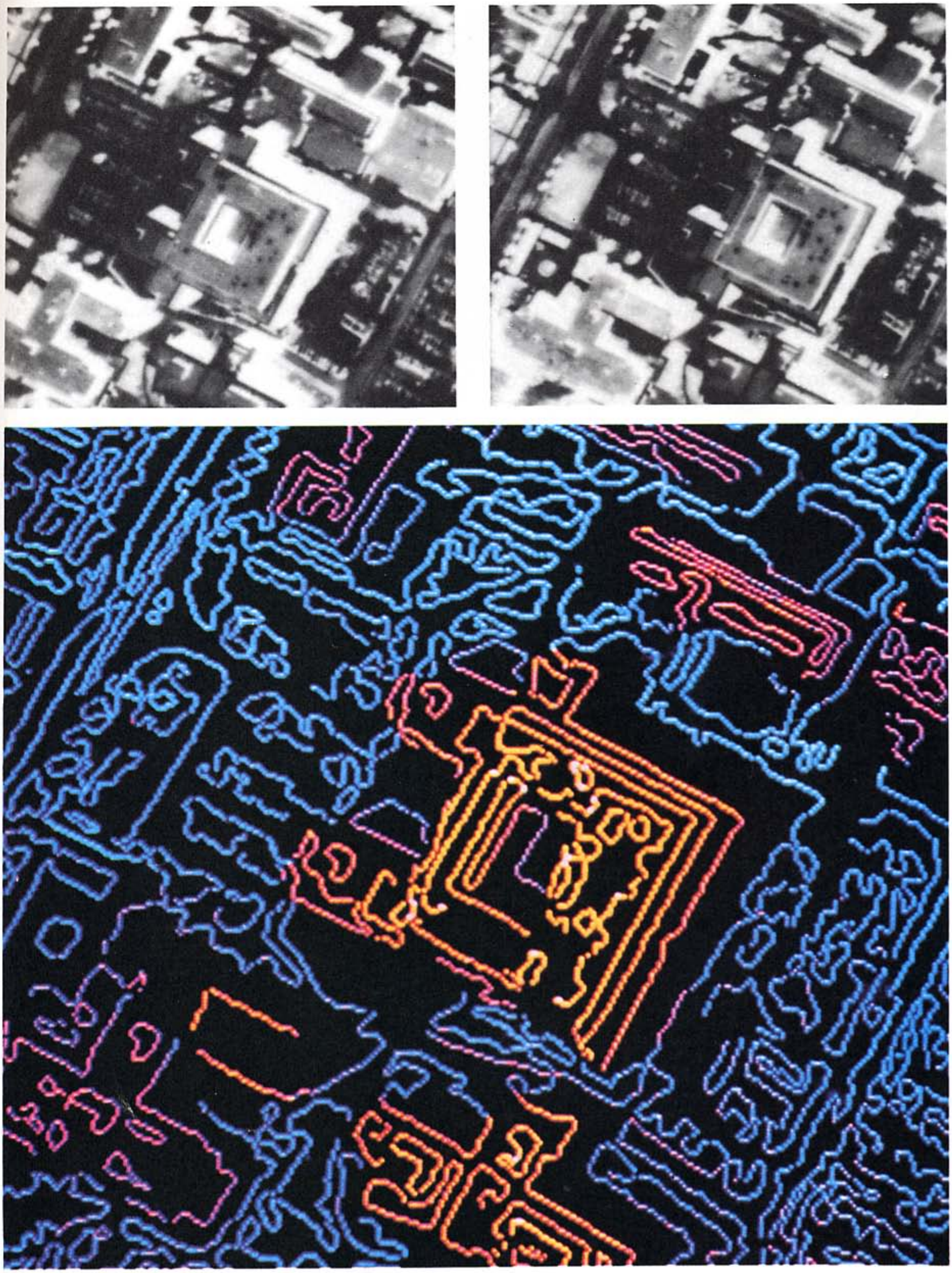
In no field have the descriptions of information-processing tasks been more precisely formulated than in the study of vision. On the one hand it is the dominant sensory modality of human beings. If we want to create robots capable of performing complex manipulative tasks in a changing environment, we must surely endow them with adequate visual powers. Yet vision remains elusive. It is something we are good at; the brain does it rapidly and easily. It is nonetheless a mammoth information-processing task. If it required a conscious effort, like adding numbers in our head, we would not undervalue its difficulty. Instead we are easily lured into oversimple, noncomputational preconceptions of what vision really entails.

Ultimately, of course, one wants to know how vision is performed by the biological hardware of neurons and their synaptic interconnections. But vision is not exclusively a problem in anatomy and physiology: how nerve cells are interconnected and how they act. From the perspective of information processing (by the brain or by a computer) it is a problem at many levels: the level of computation (What computational tasks must a visual system perform?), the level of algorithm (What sequence of steps completes the task?) and then the level of hardware (How might neurons or electronic circuits execute the algorithm?). Thus an attack on the problem of vision requires a variety of aids, including psychophysical evidence (that is, knowledge of how well people can see) and neurophysiological data (knowledge of what neurons can do). Finding workable algorithms is the most critical part of the project, because algorithms are constrained both by the computation and by the available hardware.

Figure 5.1 STEREO VISION BY A COMPUTER is shown in aerial photographs (provided by Robert J. Woodham). They were made from different angles so that objects in each have slightly different positions. The images were made by a mosaic of microelectronic sensors, each of which measures the intensity of light along a particular line of sight, as do the photoreceptor cells of the eye. The map at the bottom was generated by a computer programmed to follow a procedure devised by David Marr and the author and further developed by W. Eric L. Grimson. The computer filtered the images to emphasize spatial changes in intensity. Then it performed stereopsis: it matched features from one image to the other, determined the disparity between their positions and calculated their relative depths in the three-dimensional world. Increasing elevations in the map are coded in colors from blue to red.

Here I shall outline an effort in which I am involved, one that explores a sequence of algorithms first to extract information, notably edges, or pronounced contours in the intensity of light, from visual images and then to calculate from those edges the depths of objects in the three-dimensional world. I shall concentrate on a particular aspect of the task, namely stereopsis, or stereo vision (see Figure 5.1). Not the least of my reasons is the central role stereopsis has played in the work on vision that my colleagues and I have done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. In particular, stereopsis has stimulated a close investigation of the very first steps in visual information processing. Then too, stereopsis is deceptively simple. As with so many other tasks that the brain performs without effort, the development of an automatic system with stereo vision has proved to be surprisingly difficult. Finally, the study of stereopsis benefits from the availability of a large body of psychophysical evidence that defines and constrains the problem.

The information available at the outset of the process of vision is a two-dimensional array of measurements of the amount of light reflected into the eye or into a camera from points on the surfaces of objects in the three-dimensional visual world. In the human eye the measurements are made by photoreceptors (rod cells and cone cells), of which there are more than 100 million. In a camera that my colleagues and I use at the Artificial Intelligence Laboratory the processes are different but the result is much the same. There the measurements are made by solid-state electronic sensors. They pro-



duce an array of 1,000 by 1,000 light-intensity values. Each value is a pixel, or picture element (see Figure 5.2).

In either case it is inconceivable that the gap between the raw image (the large array of numbers produced by the eye or the camera) and vision (knowing *what* is around, and *where*) can be spanned in a single step. One concludes that vision requires various processes—one thinks of them as modules—operating in parallel on raw images and producing intermediate representations of the images on which other processes can work. For example, several vision modules seem to be involved in reconstructing the three-dimensional geometry of the world. A short list of such modules would have to include modules that deduce shape from shading, from visual texture, from motion, from contours, from occlusions and from stereopsis. Some may work directly on the raw image (the intensity measurements). Often, however, a module may operate more effectively on an intermediate representation.

Stereopsis arises from the fact that our two eyes view the visual world from slightly different angles. To put it another way, the eyes converge slightly, so that their axes of vision meet at a point in the visual world. The point is said to be fixated by the eyes, that is, the image of the point falls on the center of vision of each retina. Any neighboring point in the visual field will then project to a point on each retina some distance from the center of vision. In general this distance will not be the same for both eyes. In fact, the disparity in distance will vary with the depth of the point in the visual field with respect to the fixated point (see Figure 5.3).

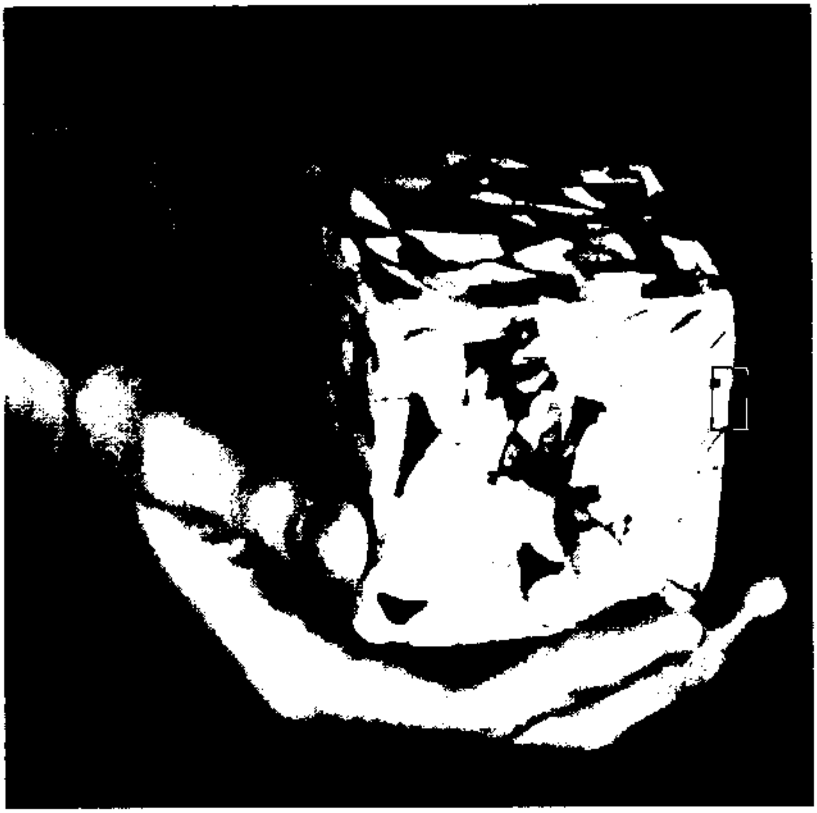
Stereopsis, then, is the decoding of three-dimensionality from binocular disparities. It might appear at first to be a straightforward problem in trigonometry. One might therefore be tempted to program a computer to solve it that way. The effort would fail; our own facility with stereopsis has led us to gloss over the central difficulty of the task, as we may see if we formally set out the steps involved in the task. They are four: A location in space must be selected from one retinal image. The same location must be identified in the other retinal image. Their positions must be measured. From the disparity between the two measurements the distance to the location must be calculated.

The last two steps are indeed an exercise in trigonometry (at least in the cases considered in this chapter). The first two steps are different. They require, in effect, that the projection of the same point in the physical world be found in each eye. A group of contiguous photoreceptors in one eye can be thought of as looking along a line of sight to a patch of the surface of some object. The photoreceptors looking at the same patch of surface from the opposite eye must then be identified. Because of binocular disparity they will not be at the same position with respect to the center of vision.

This, of course, is where the difficulty lies. For us the visual world contains surfaces that seem effectively labeled because they belong to distinct shapes in specific spatial relations to one another. One must remember, however, that vision begins with no more than arrays of raw light intensity measured from point to point. Could it be that the brain matches patterns of raw light intensity from one eye to the other? Probably not. Experiments with computers place limits on the effectiveness of the matching, and physiological and psychophysical evidence speaks against it for the human visual system. For one thing, a given patch of surface will not necessarily reflect the same intensity of light to both eyes. More important, patches of surface widely separated in the visual world may happen to have the same intensity. Matching such patches would be incorrect.

A discovery made at AT&T Bell Laboratories by Bela Julesz (now at Rutgers University) shows the full extent of the problem. Julesz devised pairs of what he called random-dot stereograms. They are visual stimuli that contain no perceptual clues except binocular disparities. To make each pair he generated a random texture of black and white dots and made two copies of it. In one of the copies he shifted an area of the pattern, say a square. In the other copy he shifted the square in the opposite direction. He filled the resulting hole in each pattern with more random texture. Viewed one at a time each pattern looked uniformly random (see Figure 5.4). Viewed through a stereoscope, so that each eye saw one of the patterns and the brain could fuse the two, the result was startling. The square gave a vivid impression of floating in front of its surroundings or behind them (see Figure 5.5). Evidently stereopsis does not require the prior perception of objects or the recognition of shapes.

Julesz' discovery enables one to formulate the computational goal of stereopsis: it is the extraction of binocular disparities from a pair of images



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

Figure 5.2 BEGINNING OF VISION for an animal or a computer is a gray-level array: a point-by-point representation of the intensity of light produced by a grid of detectors in the eye or in a digital camera. The image at the top of this illustration is such an array. It was produced by a digital camera as a set of intensity values in a grid of 576 by 454 picture elements ("pixels"). Intensity values for the part of the image inside the rectangle are given digitally at the bottom. (Figures 5.2 and 5.6 were prepared by H. Keith Nishihara of the Artificial Intelligence Laboratory.)

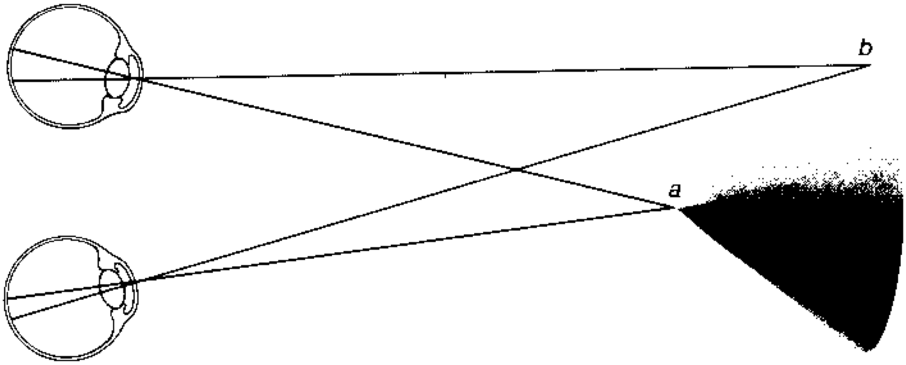


Figure 5.3 BINOCULAR DISPARITIES are the basis for stereopsis. They arise because the eyes converge slightly, so that their axes of vision meet at a point in the external world (*a*). The point is "fixated." A neighboring point in the world (*b*) will then project to a point on the retina some distance from the center of vision. The distance will not be the same for each eye.

without the need for obvious monocular clues. In addition the discovery enables one to formulate the computational problem inherent in stereopsis. It is the correspondence problem: the matching of elements in the two images that correspond to the same location in space without the recognition of objects or their parts. In random-dot stereograms the black dots in each image are all the same: they have the same size, the same shape and the same brightness. Any one of them could in principle be matched with any one of a great number of dots in the other image. And yet the brain solves this false-target dilemma: it consistently chooses only the correct set of matches.

It must use more than the dots themselves. In particular, the fact that the brain can solve the correspondence problem shows it exploits a set of implicit assumptions about the visual world, assumptions that constrain the correspondence problem, making it determined and solvable. In 1976 David Marr and I, working at MIT, found that simple properties of physical surfaces could limit the problem sufficiently for the stereopsis algorithms (procedures to be followed by a computer) we were then investigating. These are, first, that a given point on a physical surface has only one three-dimensional location at any given time and, second, that physical objects are cohesive and usually are opaque, so that the variation in depth over a surface is generally smooth, with discontinuous changes occurring only at boundary lines. The first of these constraints—uniqueness of location—means that each item in either image (say each dot in a random-dot stereogram) has a unique disparity and can be matched with no more than one item in the other image. The second constraint—continuity and opacity—means that disparity varies smoothly except at object boundaries.

Together the two constraints provide matching rules that are reasonable and powerful. I shall describe some simple ones below. Before that, however, it is necessary to specify the items to be matched. After all, the visual world is not a random-dot stereogram, consisting only of black and white dots. We have already seen that intensity values are too unreliable. Yet the information the brain requires is encrypted in the intensity array provided by photoreceptors. If an additional property of physical surfaces is invoked, the problem is simplified. It is based on the observation that at places where there are physical changes in a surface, the image of the surface usually shows sharp variations in intensity. These variations (caused by markings on a surface and by variations in its depth) would be more reliable tokens for matching than raw intensities would be.

Instead of raw numerical values of intensity, therefore, one seeks a more symbolic, compact and robust representation of the visual world: a description of the world in which the primitive symbols—the signs in which the visual world is coded—are intensity variations. Marr called it a "primal sketch." In essence it is the conversion of the gray-level arrays provided by the visual photoreceptors

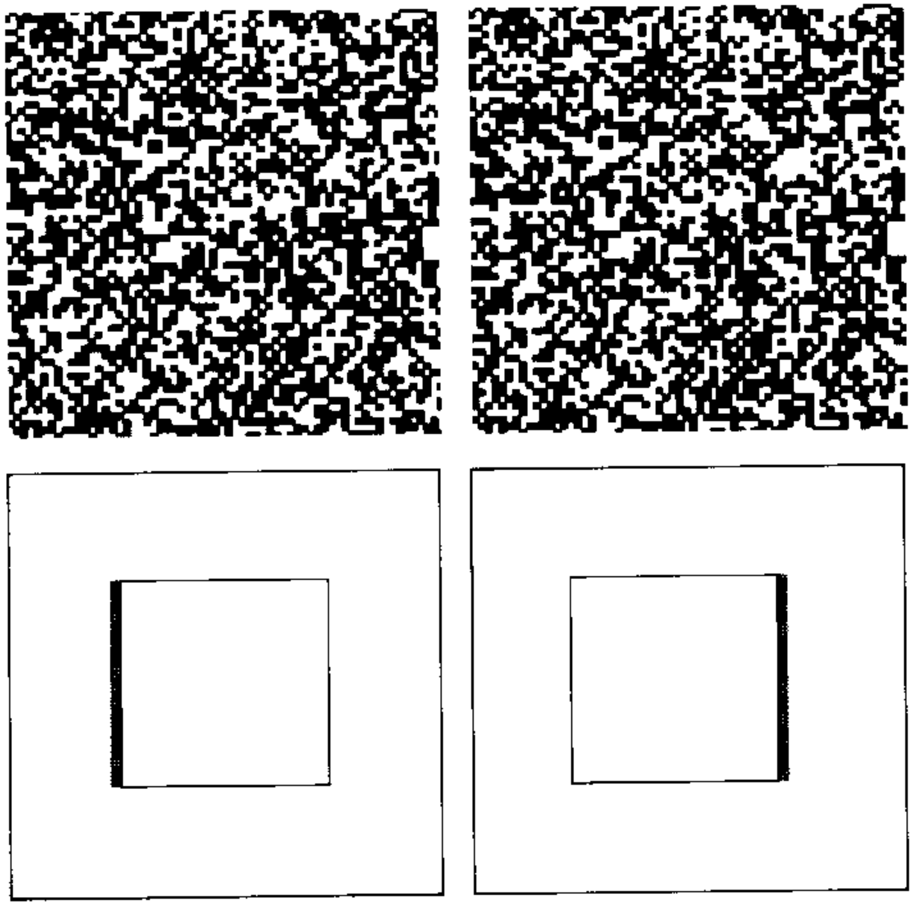


Figure 5.4 RANDOM-DOT STEREOGRAMS devised by Bela Julesz working at AT&T Bell Laboratories are visual textures containing no clues for stereo vision except binocular disparities. The stereograms themselves are the same random texture of black and white dots (*top*). In one of them, however, a square of the texture is shifted toward the left; in the other it is shifted toward the right (*bottom*). The resulting hole in each image is filled with more random dots (*gray areas*).

into a form that makes explicit the position, direction, scale and magnitude of significant light-intensity gradients, with which the brain's stereopsis module can solve the correspondence problem and reconstruct the three-dimensional geometry of the visual world. I shall describe a scheme we have been using at the Artificial Intelligence Laboratory for the past few years, based on old and new ideas developed by a number of investigators, primarily Marr, Ellen C. Hildreth and me. It has several attractive features: it is fairly simple, it works well and it shows interesting resemblances to biological vision, which, in fact, suggested it. It is not, however, the full solution. Perhaps it is best seen as a working hypothesis about vision.

Basically the changes of intensity in an image can be detected by comparing neighboring intensity values in the image: if the difference between them is great, the intensity is changing rapidly. In mathematical terms the operation amounts to taking the first derivative. (The first derivative is simply the rate of change of a mathematical function. Here it is simply the rate at which intensity changes on a path across the gray-level array.) The position of an extremal value—a peak or a pit—in the first deriva-

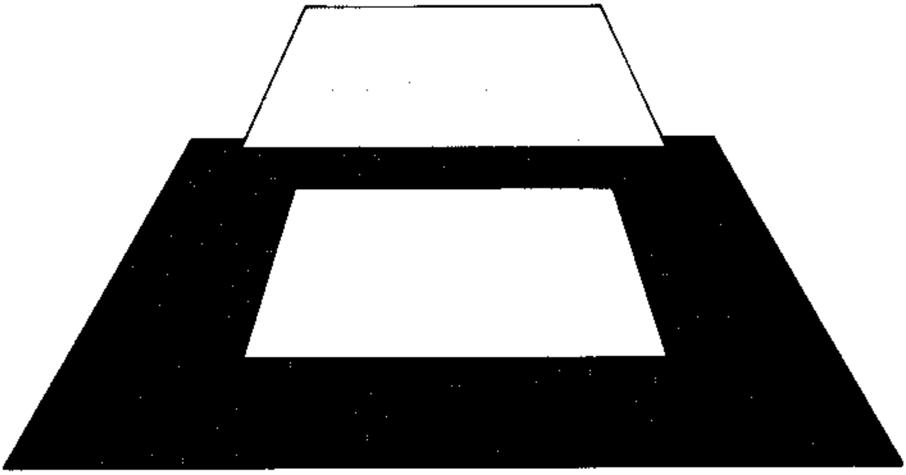


Figure 5.5 VIVID PERCEPTION OF DEPTH results when the random-dot stereograms shown in Figure 5.4 are viewed through a stereoscope, so that each eye sees one of the pair and the brain can fuse the two. The sight of part of the image "floating" establishes that stereopsis does not require the recognition of objects in the visual world.

tive turns out to localize the position of an intensity edge quite well (see Figure 5.6) In turn the intensity edge often corresponds to an edge on a physical surface. The second derivative also serves well. It is simply the rate of change of the rate of change and is obtained by taking differences between neighboring values of the first derivative. In the second derivative an intensity edge in the gray-level array corresponds to a zero-crossing: a place where the second derivative crosses zero as it falls from positive values to negative values or rises from negative values to positive.

Derivatives seem quite promising. Used alone, however, they seldom work on a real image, largely because the intensity changes in a real image are rarely clean and sharp changes from one intensity value to another. For one thing, many different changes, slow and fast, often overlap on a variety of different spatial scales. In addition changes in intensity are often corrupted by the visual analogue of noise. They are corrupted, in other words, by random disruptions that infiltrate at different stages as the image formed by the optics of the eye or of a camera is transduced into an array of intensity measurements. In order to cope both with noisy edges and with edges at different spatial scales the image must be "smoothed" by a local averaging of neighboring intensity values. The differencing operation that amounts to the taking of first and second derivatives can then be performed.

There are various ways the sequence can be managed, and much theoretical effort has gone into the search for optimal methods. In one of the simplest the two operations—smoothing and differentiation—are combined into one. In technical terms it sounds forbidding: the image is convolved with a filter that embodies a particular center-surround function, the Laplacian of a Gaussian. It is not as bad as it sounds. A two-dimensional Gaussian is the bell-shaped distribution familiar to statisticians. In this context it specifies the importance to be assigned to the neighborhood of each pixel when the image is being smoothed. As the distance increases, the importance decreases. A Laplacian is a second derivative that gives equal weight to all paths extending away from a point. The Laplacian of a Gaussian converts the bell-shaped distribution into something more like a Mexican hat. The bell is narrowed and at its sides a circular negative dip develops.

Now the procedure can be described nontechnically. Convolving an image with a filter that embodies the Laplacian of a Gaussian is equivalent to substituting for each pixel in the image a weighted average of neighboring pixels, where the weights are provided by the Laplacian of a Gaussian. Thus the filter is applied to each pixel. It assigns the greatest positive weight to that pixel and decreasing positive weights to the pixels nearby (see Figure 5.7). Then comes an annulus—a ring—in

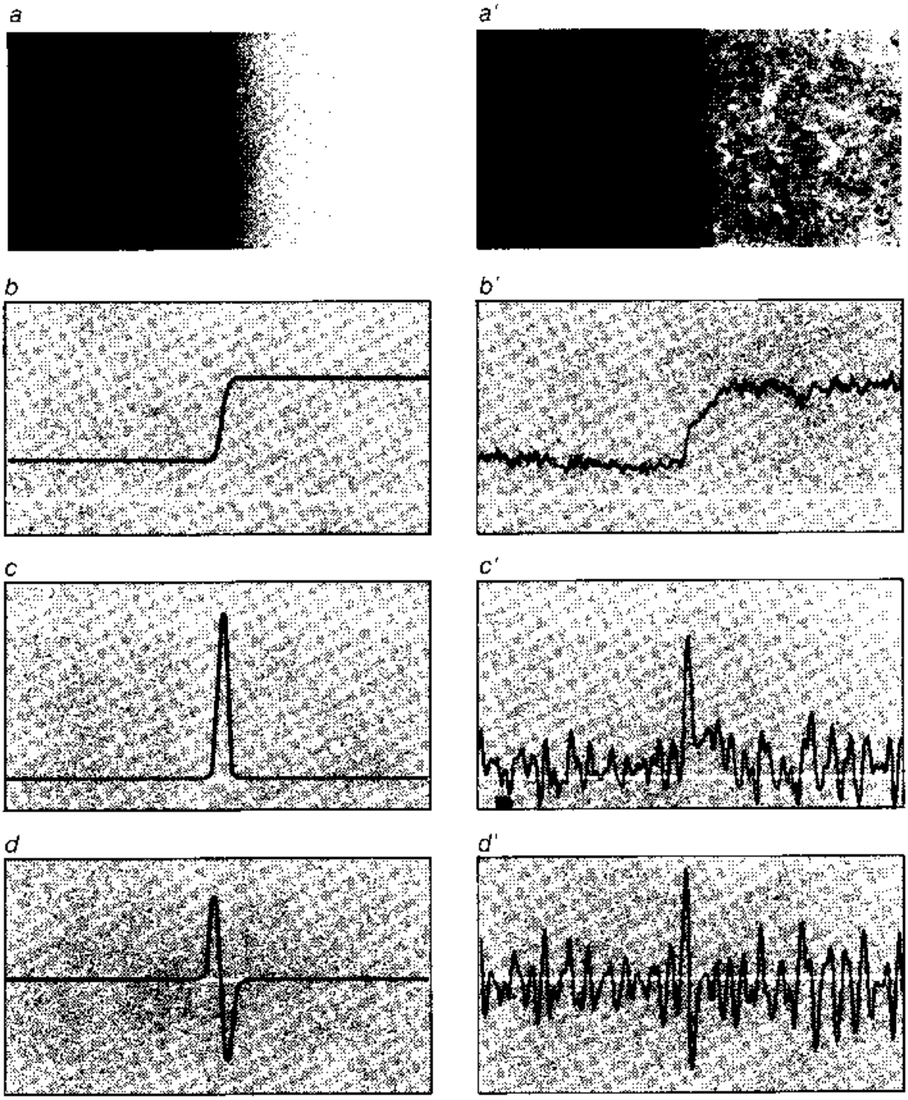


Figure 5.6 SPATIAL DERIVATIVES of an image emphasize its spatial variations in intensity. Left: An edge is shown between two even shades of gray (*a*). The intensity along a path across the edge appears below it (*b*). The first derivative of the intensity is the rate at which intensity changes (*c*). Toward the left or toward the right there is no change; the first derivative therefore is zero. Along the edge itself, however, the rate of change rises and falls. The second derivative of the intensity is the rate of change of the rate of change (*d*). Both derivatives emphasize the edge. The first derivative marks it with a peak; the second derivative marks it by crossing zero. Right: The edge is more typical of the visual world (*a′*). The related intensity contour (*b′*) and its first and second derivatives (*c′*, *d′*) are "noisy." The edge must be smoothed before derivatives are taken.

which the pixels are given negative weightings. Bright points there feed negative numbers into the averaging. The result of the overall filtering is an array of positive and negative numbers: a kind of second derivative of the image intensity at the scale of the filter. The zero-crossings in this filtered array correspond to places in the original image where its intensity changes most rapidly. Note that a binary (that is, a two-valued) map showing merely the positive and negative regions of the filtered array is

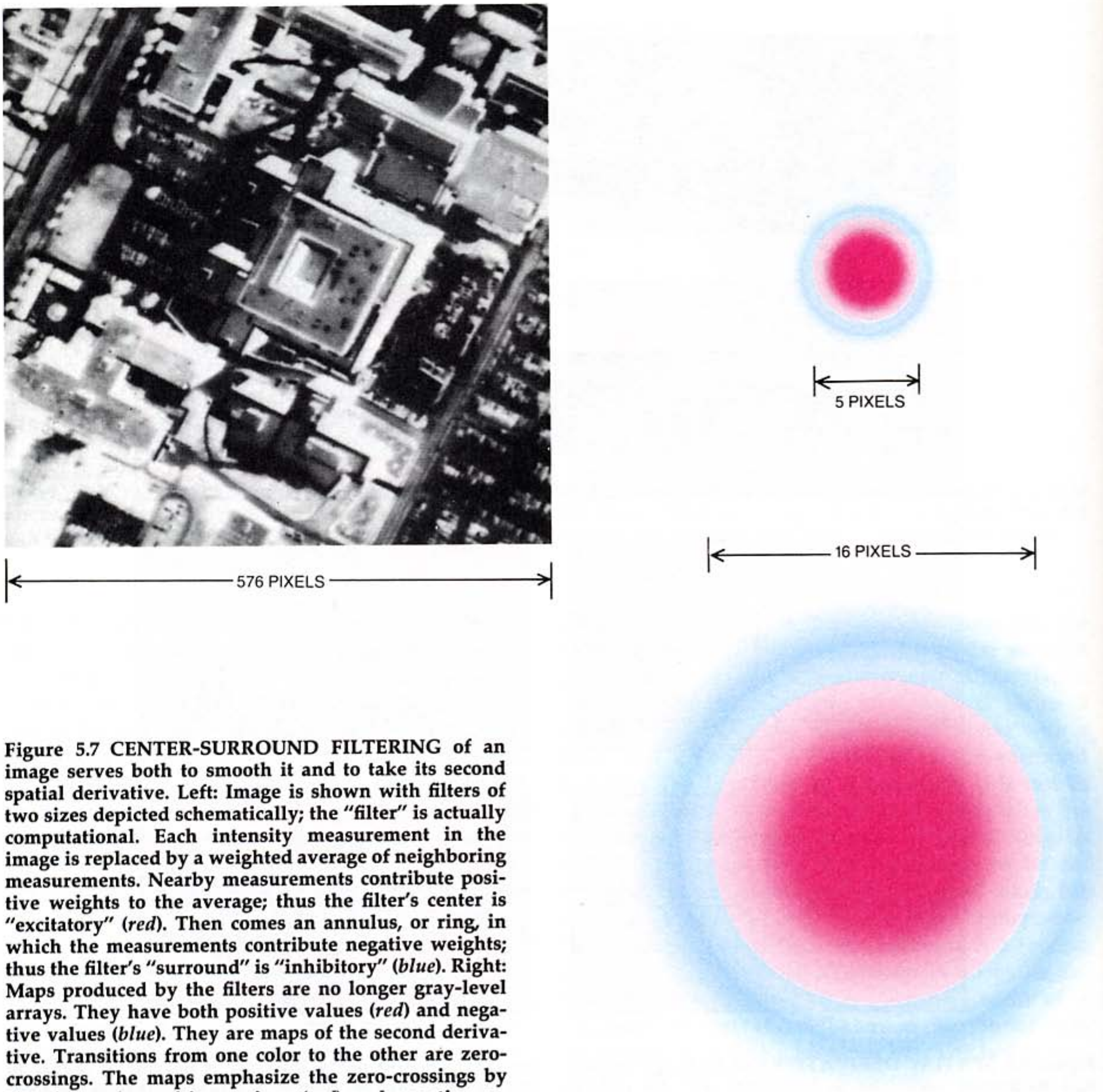
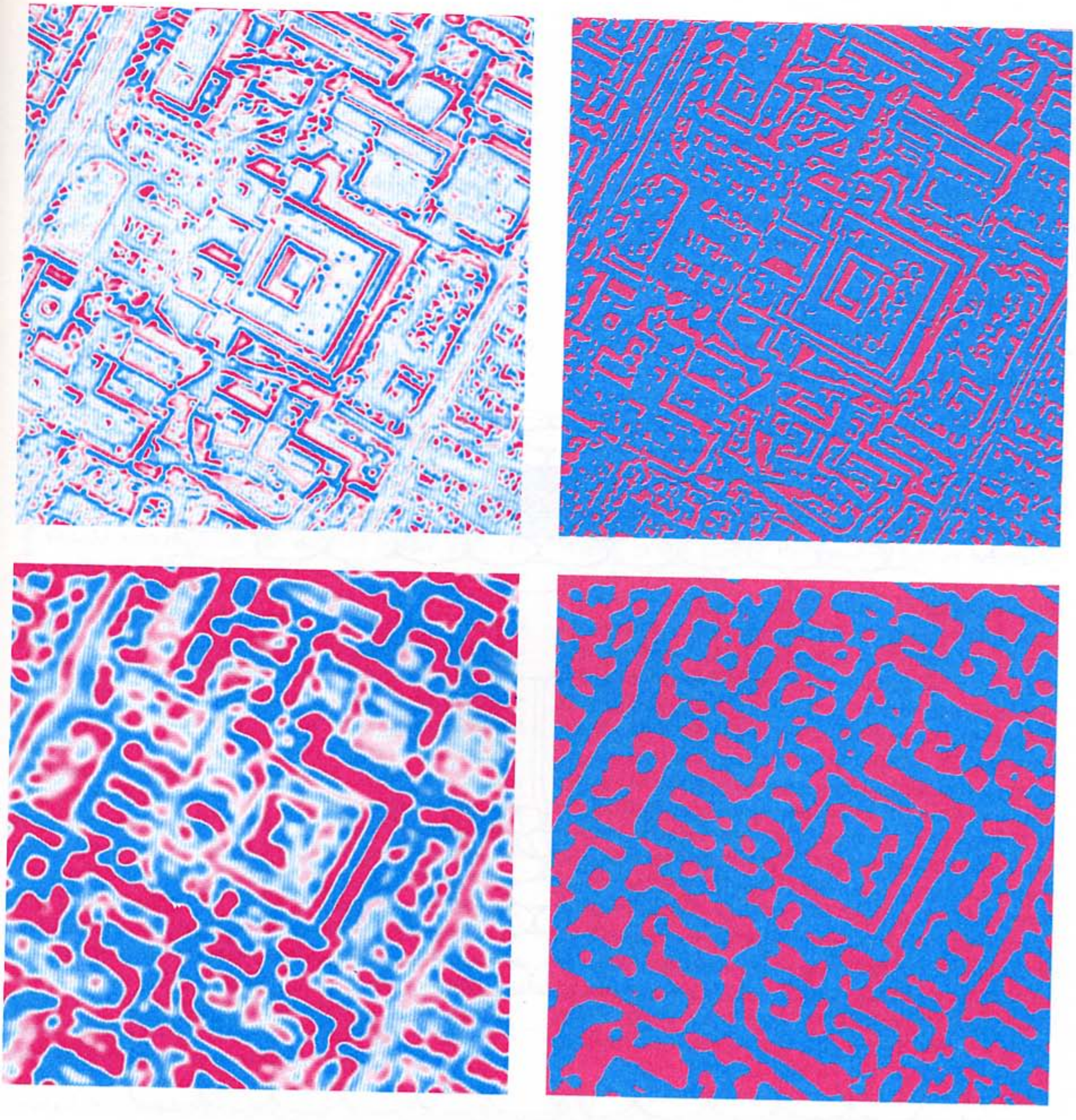


Figure 5.7 CENTER-SURROUND FILTERING of an image serves both to smooth it and to take its second spatial derivative. Left: Image is shown with filters of two sizes depicted schematically; the "filter" is actually computational. Each intensity measurement in the image is replaced by a weighted average of neighboring measurements. Nearby measurements contribute positive weights to the average; thus the filter's center is "excitatory" (*red*). Then comes an annulus, or ring, in which the measurements contribute negative weights; thus the filter's "surround" is "inhibitory" (*blue*). Right: Maps produced by the filters are no longer gray-level arrays. They have both positive values (*red*) and negative values (*blue*). They are maps of the second derivative. Transitions from one color to the other are zero-crossings. The maps emphasize the zero-crossings by showing only positive regions (*red*) and negative regions (*blue*).

essentially equivalent to a map of the zero-crossings in that one can be constructed from the other.

In the human brain most of the hardware required to perform such a filtering seems to be present. As early as 1865 Ernst Mach observed that visual perception seems to enhance spatial variations in light intensity. He postulated that the enhancement might be achieved by lateral inhibition, a brain mechanism in which the excitation of an axon, or nerve fiber, say by a spot of bright light in the visual world, blocks the excitation of neighboring axons. The operation plainly enhances the con-



trast between the bright spot and its surroundings. Hence it is similar to the taking of a spatial derivative.

Then in the 1950's and 1960's evidence accumulated suggesting that the retina does something much like center-surround filtering. The output from each retina is conveyed to the rest of the brain by about a million nerve fibers, each being the axon of a neuron called a retinal ganglion cell. The cell derives its input (by way of intermediate neurons) from a group of photoreceptors, which form a "receptive field." What the evidence suggests is that for certain ganglion cells the receptive field has a center-surround organization closely approximating the Laplacian of a Gaussian. Brightness in the center of the receptive field excites the ganglion cell; brightness in a surrounding annulus inhibits it. In short, the receptive field has an ON-center and an OFF-surround, just like the Mexican hat (see Figure 5.8).

Other ganglion cells have the opposite properties: they are OFF-center, ON-surround. If axons could signal negative numbers, these cells would be redundant: they report simply the negation of what the ON-center cells report. Neurons, however, cannot readily transmit negative activity; the ones that transmit all-or-nothing activity are either active or quiescent. Nature, then, may need neuronal opposites. Positive values in an image subjected to center-surround filtering could be represented by the activity of ON-center cells; negative values could be represented by the activity of OFF-center cells. In this regard I cannot refrain from mentioning the recent finding that ON-center and OFF-center ganglion cells are segregated into two different layers, at least in the retina of the cat. The maps generated

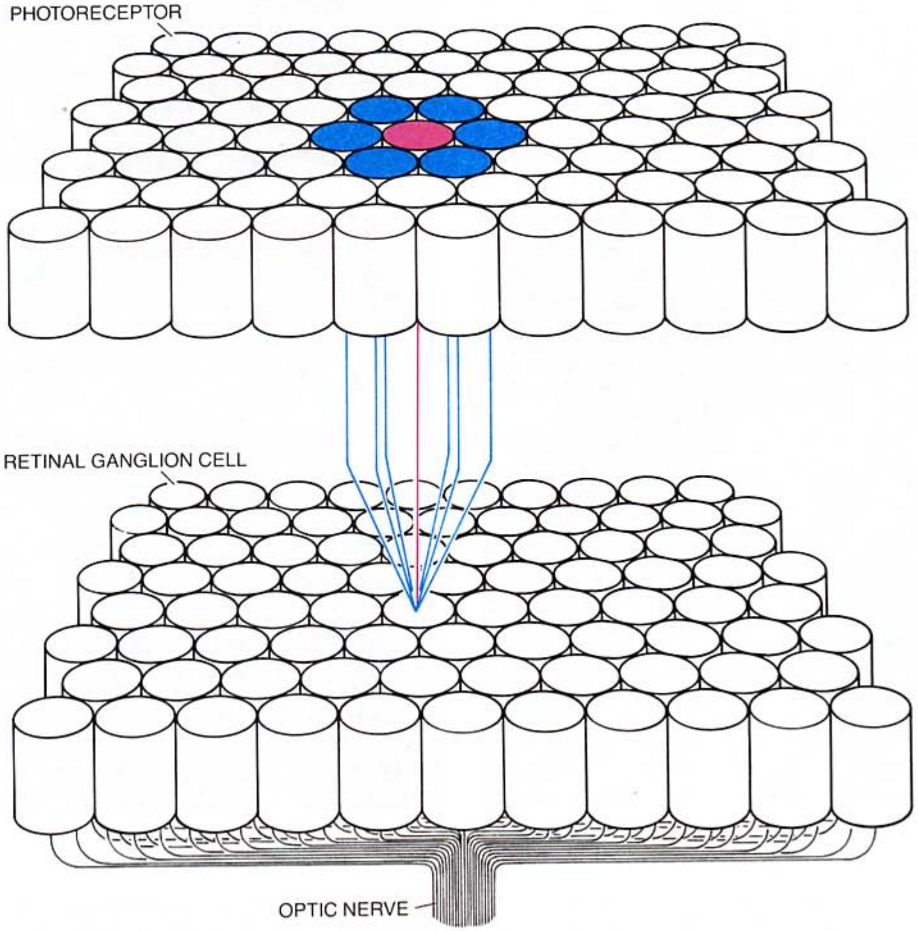


Figure 5.8 BIOLOGICAL CENTER-SURROUND FILTER embodied by cells in the retina resembles the computer procedure shown in Figure 5.7. The filter begins with a layer of photoreceptors connected by way of intermediate nerve cells, not shown in the diagram, to a layer of retinal ganglion cells, which send visual data to higher visual centers. For the sake of simplicity only one set of connections is shown. A photoreceptor cell (*red*) excites an "ON-center" ganglion cell by promoting its tendency to generate neural signals; the surrounding photoreceptors (*blue*) inhibit the ganglion cell.

by our computer might thus depict neural activity rather literally. In the maps in Figure 5.7 red might correspond to ON-layer activity and blue to OFF-layer activity. Zero-crossings (that is, transitions from one color to the other) would be the locations where activity switches from one layer to the other. Here, then, is a conjecture linking a computational theory of vision to the brain hardware serving biological vision.

It should be said that the center-surround filtering of an image is computationally expensive for a computer because it involves great numbers of multiplications: about a billion for an image of 1,000 pixels by 1,000. At the Artificial Intelligence Laboratory, H. Keith Nishihara and Noble G. Larson, Jr., have designed a specialized device: a convolver that performs the operation in about a second. The speed is impressive but is plodding compared with that of the retinal ganglion cells.

I should also mention the issue of spatial scale. In an image there are fine changes in intensity as well as coarse. All must be detected and represented. How can it be done? The natural solution (and the solution suggested by physiology and psychophysics) is to use center-surround filters of different sizes. The filters turn out to be band-pass: they respond optimally to a certain range of spatial frequencies. In other words, they "see" only changes in intensity from pixel to pixel that are neither too fast nor too slow. For any one spatial scale the process of finding intensity changes consists, therefore, of filtering the image with a center-surround filter (or receptive field) of a particular size and then finding the zero-crossings in the filtered image. For a combination of scales it is necessary to set up filters of different sizes, performing the same computation for each scale. Large filters would then detect soft or blurred edges as well as overall illumination changes; small filters would detect finer details. Sharp edges would be detected at all scales.

Recent theoretical results enhance the attractiveness of this idea by showing that features similar to zero-crossings in a filtered image can be rich in information. First, Ben Logan of Bell Laboratories has proved that a one-dimensional signal filtered through a certain class of filters can be reconstructed from its zero-crossings alone. The Laplacian of a Gaussian does not satisfy Logan's conditions exactly. Still, his work suggests that the primitive symbols provided by zero-crossings are potent visual symbols. More recently Alan Yuille and I have made a theoretical analysis of center-surround filtering. We have been able to show that zero-crossing maps obtained at different scales can represent the original image completely, that is, without any loss of information.

This is not to say that zero-crossings are the optimal coding scheme for a process such as stereopsis. Nor is it to insist that zero-crossings are the sole basis of biological vision. They are a candidate for an optimal coding scheme, and they (or something like them) may be important among the items to be matched between the two retinal images. We have, therefore, a possible answer to the question of what the stereopsis module matches. In addition we have the beginning of a computational theory that may eventually give mathematical precision to the vague concept of "edges" and connect it to known properties of biological vision, such as the prominence of "edge detector" cells discovered at the Harvard Medical School by David H. Hubel and Torsten N. Wiesel in the part of the cerebral cortex where visual data arrive.

To summarize, a combination of computational arguments and biological data suggests that an important first step for stereopsis and other visual processes is the detection and marking of changes in intensity in an image at different spatial scales. One way to do it is to filter the image by the Laplacian of a Gaussian; the zero-crossings in the filtered array will then correspond to intensity edges in the image. Similar information is implicit in the activity of ON-center and OFF-center ganglion cells in the retina. To explicitly represent the zero-crossings (if indeed the brain does it at all) a class of edge-detector neurons in the brain (no doubt in the cerebral cortex) would have to perform specific operations on the output of ON-center and OFF-center cells that are neighbors in the retina. Here, however, one comes up against the lack of information about precisely what elementary computations nerve cells can readily do.

We are now in a position to see how a representation of intensity changes might be useful for stereopsis. Consider first an algorithm devised by Marr and me that implements the constraints discussed above, namely uniqueness (a given point on a physical surface has only one location, so that only one binocular match is correct) and continuity (variations in depth are generally smooth, so that binocular disparities tend to vary smoothly). It is successful at solving random-dot stereograms and at least some natural images. It is done by a computer; thus its actual execution amounts to a sequence of calculations. It can be thought of, however, as setting up a three-dimensional network of nodes, where the nodes represent all possible intersections of lines of sight from the eyes in the three-dimensional world. The uniqueness constraint will then be implemented by requiring that the nodes along a given line of sight inhibit one another. Meanwhile the continuity constraint will be implemented by requiring that each node excite its neighbors. In the case of random-dot stereograms the procedure will be relatively simple. There the matches for pixels on each horizontal row in one stereogram need be sought only along the corresponding row of the other stereogram.

The algorithm starts by assigning a value of 1 to all nodes representing a binocular match between two white pixels or two black pixels in the pair of stereograms. The other nodes are given a value of 0. The 1's thus mark all matches, true and false (see Figure 5.9). Next the algorithm performs an algebraic sum for each node. In it the neighboring nodes with a value of 1 contribute positive weights; the nodes with a value of 1 along lines of sight contribute negative weights. If the result exceeds some threshold value, the node is given the value of 1; otherwise the node is set to 0. That constitutes one iteration of the procedure. After a few such iterations the network reaches stability. The stereopsis problem is solved (see Figure 5-10).

The algorithm has some great virtues. It is a cooperative algorithm: it consists of local calculations that a large number of simple processors could perform asynchronously and in parallel. One imagines that neurons could do them. In addition the algorithm can fill in gaps in the data. That is, it interpolates continuous surfaces. At the same time it allows for sharp discontinuities. On the other hand, the network it would require to process finely detailed natural images would have to be quite large, and most of the nodes in the network would be idle at any one time. Furthermore, intensity values are unsatisfactory for images more natural than random-dot stereograms.

The algorithm's effectiveness can be extended to at least some natural images by first filtering the images to obtain the sign of their convolution with the Laplacian of a Gaussian. The resulting binary maps then serve as inputs for the cooperative algorithm. The maps themselves are intriguing. In the ones generated by large filters at correspondingly low spatial resolution, zero-crossings of a given sign (for instance the crossings at which the sign of the convolution changes from positive to negative) turn out to be quite rare and are never close to each other. Thus false targets (matches between noncorresponding zero-crossings in a pair of stereograms) are essentially absent over a large range of disparities.

This suggests a different class of stereopsis algorithms. One such algorithm, developed recently for robots by Nishihara, matches positive or negative patches in filtered image pairs. Another algorithm, developed earlier by Marr and me, matches zero-crossings of the same sign in image pairs made by filters of three or more sizes. First the coarsely filtered images are matched and the binocular disparities are measured. The results are employed to approximately register the images. (Monocular features such as textures could also be used.) A similar matching process is then applied to the medium-filtered images. Finally the process is applied to the most finely filtered images. By that time the binocular disparities in the stereo pair are known in detail, and so the problem of stereopsis has been reduced to trigonometry.

A theoretical extension and computer implementation of our algorithm by W. Eric L. Grimson at the Artificial Intelligence Laboratory works quite well for a typical application of stereo systems: the analysis of aerial photographs (see Figure 5.1). In addition it mimics many of the properties of human depth perception. For example, it performs successfully when one of the stereo images is out of focus. Yet there may also be subtle differences. Recent work by John Mayhew and John P. Frisby at the University of Sheffield and by Julesz at Bell Laboratories should clarify the matter.

What can one say about biological stereopsis? The algorithms I have described are still far from solving the correspondence problem as effectively as our own brain can. Yet they do suggest how the problem is solved. Meanwhile investigations of the cerebral cortex of the cat and of the cerebral cortex of the macaque monkey have shown that certain cortical neurons signal binocular disparities. And quite recently Gian F. Poggio of the Johns Hopkins University School of Medicine has found cortical neurons that signal the correct binocular disparity in random-dot stereograms in which there are many false matches. His discovery, together with our computational analysis of stereopsis, promises to yield insight into the brain mechanisms underlying depth perception.

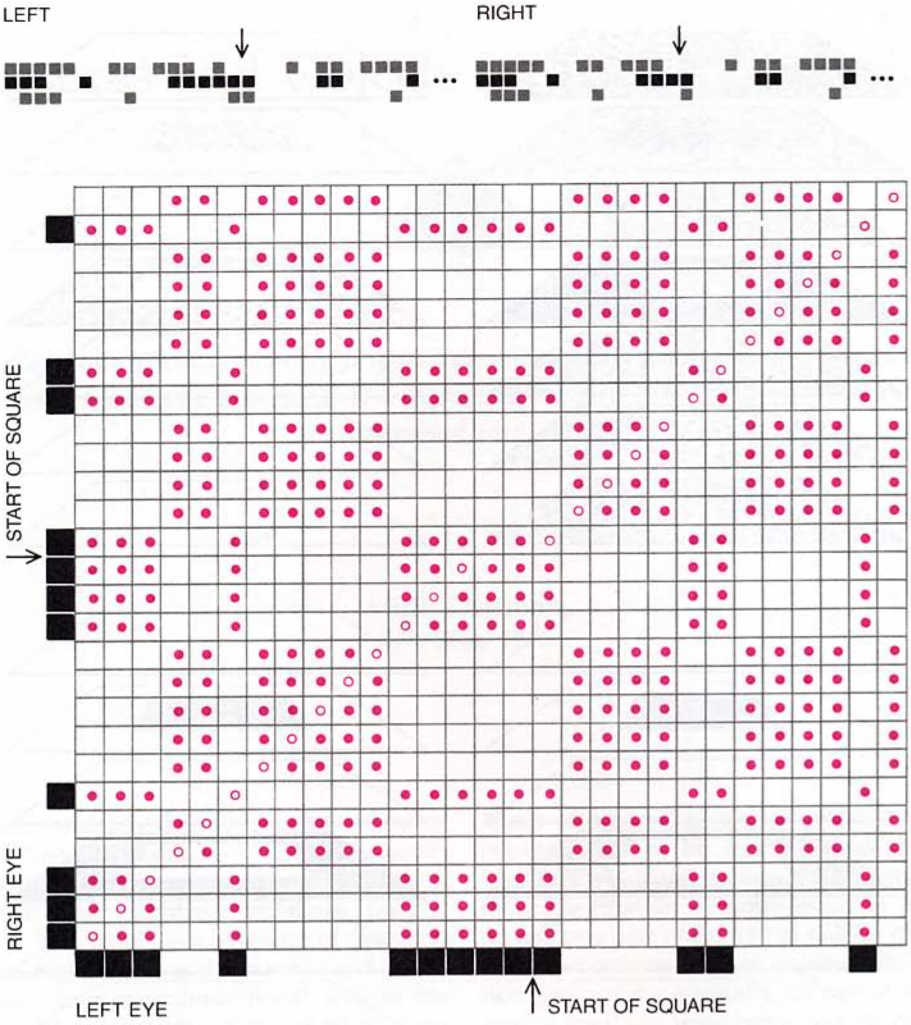


Figure 5.9 STEREOPSIS ALGORITHM reconstructs the three-dimensional visual world by seeking matches between dots on corresponding rows of a pair of random-dot stereograms. Two such rows are shown (*black and white, top*). The rows below are placed along the axes. Horizontal lines represent lines of sight for the right eye; vertical lines, lines of sight for the left. Color marks all intersections at which the eyes both see a black dot or a white dot. A given dot in one stereogram could in principle match any same-color dot in the other. Yet only some matches are correct (*open colored circles*), that is, only some reveal that a square of random-dot texture has a binocular disparity.

One message should emerge clearly: the extent to which the computer and the brain can be brought together for the study of problems such as vision. On the one hand the computer provides a powerful tool for testing computational theories and algorithms. In the process it guides the design of neurophysiological experiments: it suggests what one should look for in the brain. The impetus this will give brain research in the coming decades is likely to be great.

The benefit is not entirely in that direction; computer science also stands to gain. Some computer scientists have maintained that the brain provides only existence proofs, that is, a living demonstration that a given problem has a solution. They are mistaken. The brain can do more: it can show how to seek solutions. The brain is an information processor that has evolved over many millions of years to perform certain tasks superlatively well. If we regard it, with justified modesty, as an uncertain in-

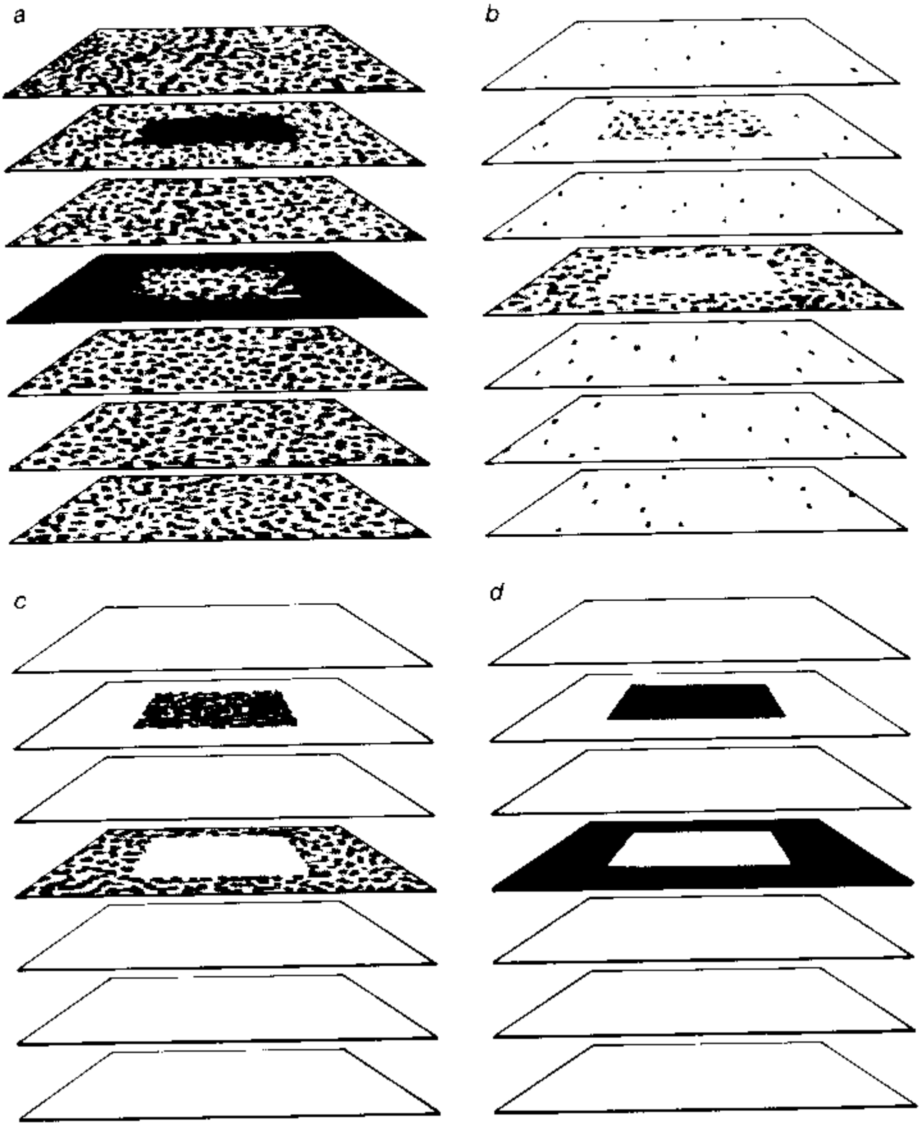


Figure 5.10 ITERATIONS OF THE ALGORITHM (depicted schematically) solve the problem of stereopsis. The algorithm assigns a value of 1 to all intersections of lines of sight marked by a match, and of 0 to the others. Next it calculates a weighted sum for every intersection. Neighboring intersections with a value of 1 contribute positive weights to the sum. The eye sees only one surface along a given line of sight; hence intersections with a value of 1 along lines of sight contribute negative weights. If the result exceeds a threshold value, the intersection is reset to 1; otherwise it is reset to 0. After a few iterations of the procedure the calculation is complete: the stereograms are decoded.

strument, the reason is simply that we tend to be most conscious of the things it does least well—the recent things in evolutionary history, such as logic, mathematics and philosophy—and that we tend to be quite unconscious of its true powers, say in vision. It is in the latter domains that we have much to learn from the brain, and it is in these domains that we should judge our achievements in computer science and in robots. We may then begin to see what vast potential lies ahead.