MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A. I. Memo 776 March, 1984

Vision by Man and Machine:

How the brain processes visual information may be suggested by studies in computer vision (and vice versa)

Tomaso Poggio

Abstract: The development of increasingly sophisticated and powerful computers in the last few decades has frequently stimulated comparisons between them and the human brain. Such comparisons will become more earnest as computers are applied more and more to tasks formerly associated with essentially human activities and capabilities. The expectation of a coming generation of 'intelligent' computers and robots with sensory, motor and even 'intellectual' skills comparable in quality to (and quantitatively surpassing) our own is becoming more widespread and is, I believe, leading to a new and potentially productive analytical science of 'information processing'.

In no field has this new approach been so precisely formulated and so thoroughly exemplified as in the field of vision. As the dominant sensory modality of man, vision is one of the major keys to our mastery of the environment, to our understanding and control of the objects which surround us. If we wish to create robots capable of performing complex manipulative tasks in a changing environment, we must surely endow them with (among other things) adequate visual powers. How can we set about designing such flexible and adaptive robots? In designing them, can we make use of our rapidly growing knowledge of the human brain, and if so, how at the same time, can our experience in designing artificial vision systems help us to understand how the brain analyzes visual information?

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under the Office of Naval Research contract N00014-80-C-0505. This is one of the original versions of a paper which appeared in *Scientific American*, April 1984, where figures can be found.

The Information Processing Approach to Vision

When we compare brains with computers we initially find many more differences than similarities. The differences are most obvious at the level of hardware: the nerve cells, or neurons, are small, delicately shaped structures consisting of a thin and complex membrane, closely packed in a supporting medium of glial 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 organization the differences are also great. In the brain the interconnections between neurons are very numerous—each neuron may receive many thousands of inputs— and are distributed in three dimensions. In computers the wires linking the circuit components are limited by present day solid state technology to a relatively small number arranged more or less two dimensionally. The modes of transmission of signals are dissimilar as well. The binary coded electrical pulses of the computer are to some extent mirrored in the all-or-nothing action potentials of the nerve axon, but in addition to these the brain uses graded potentials, localized chemical transmission across synapses, and active ionic channels in the cells' membrane that can be modulated by a range of chemical transmitter substances. The brain does not seem to have anything similar to the registers and binary addresses of a conventional computer. Finally, the temporal organization of the two systems is dramatically different. Computers process information serially at a very fast rate. Their time course is divided up by a system-wide clock into a series of temporally discrete states. The transient changes from one state to the next play no part in the functioning of the system. What we know of brains, on the other hand, points to their functioning as much slower, parallel processors, analyzing information along millions of channels concurrently and with no regular clock-driven operations.

Given this catalogue of differences at the lower levels of hardware and system organization, at what level or levels can we begin to find similarities? Clearly, there is a level at which any two functionally operational systems can be considered equivalent, namely the level of the description of the task they perform. "To bring the good news from Ghent to Aixe" is a description of a task that can be performed by telegraph, satellite, horseback messenger or pigeon post equally well (unless other constraints such as time are specified). If, therefore, we assert that brains and computers are both functioning as information processing systems capable of performing similar tasks, we can develop descriptions or formulations of these tasks that are equally applicable to both. We have a common language with which to discuss them — the language of information processing. This independence or separability of the description of the task to be performed from the hardware performing it is a familiar notion in everyday life, but it is not always applied systematically in science. This principle is at the foundations of the new field of Artificial Intelligence, whose central goals are to make computers more useful by endowing them with more "intelligent" capabilities, and to understand the principles that make intelligence possible.

Ultimately, we would like to understand the means by which a task such as vision is performed by the biological hardware of neurons and synapses. But vision is not only a problem in physiology and anatomy — how cells are connected and behave — but also a problem in information processing. Both from the perspectives of the neurosciences and of Artificial Intelligence,a satisfactory explanation of vision requires different levels of understanding,from the level of the information processing task at hand to the level of the particular algorithm, one sequence of steps that a system executes to perform a task, to and the level of the computer or biological hardware. The power of the approach lies in the integration of all the levels of analysis; and in practice information from all sources and levels is needed to attack a particular problem. The information processing approach to vision is only at its beginnings and we are still far from an integrated understanding. A major recent contribution is David Marr's outline of how the visual perception of objects might evolve (Vision, Freeman). But despite its originality, it is far from being adequate as a theory of vision.

In general terms the goal of vision is to tell us, or a machine, *what is in the world and where*. This information is crucial for moving around without bumping into obstacles, and for describing, recognizing, localizing and manipulating objects. Problems of the class we are dealing with here are especially elusive to the intuition. Visual perception for example, is something at which we are very good at but which is not open to introspective examination. It is a mammoth information processing task which the brain performs rapidly and easily – effortlessly. If some conscious mental computation were required, as when we do mental arithmetic for example, we would not underestimate its difficulty. Instead, we are too easily lured into oversimple, non-computational preconceptions of what vision entails.

Vision begins with a large array of measurements of the amount of light that is reflected from physical surfaces in the environment onto the eye or a camera. In the human eye these measurements are made by over one million that undergo chemical changes in response to light. The cameras we use at the MIT Artificial Intelligence Laboratory measure light intensity with a physical array of so-called charge-coupled sensors. Although quite different physical processes are used, both the eye and a camera produce visual images that can be thought of as a large array of numbers that represent the intensity of light at different locations in the world.

From this array of numbers, it is too difficult to achieve in one single step an understanding of what is seen. Visual perception must be accomplished in stages by several different visual processes operating in parallel to produce intermediate *representations* of visual information, that successively approach our final understanding of the world around us. At each stage, these representations make explicit information that is possible from previous representations, and to subsequent processes. One of the final and most important representations that is useful for recognizing objects is a description of the shapes of objects around the viewer but that is independent of its viewpoint. It can be argued – as D. Marr forcefully did – that the construction of such a description requires an intermediate representation of the visible surfaces around the viewer – their distance and orientation – from the vantage point of the viewer. In turn, there are several earlier visual processes that can construct this intermediate representation from the 2D intensity arrays in which this information is encoded by the imaging process. These processes exploit visual cues available in shading, occlusion, contour, texture, motion and stereopsis. Although some of these processes may work directly on the raw intensity image, they often operate more effectively on another intermediate representation of primitive features in the image, such as edges. In the rest of the article I will describe the process of stereoscopic vision, and how the image must first be transformed into this more compact representation, on which the stereo process can then operate.

Two notes of caution are needed here. First, while it seems that any powerful vision system must use several intermediate representations, it is not yet clear how the different representations interact and how the system as a whole is organized. Most likely, the flow of information is not simply *bottom up* through the various representations, with each process operating only on the immediately preceding representation. Second, different processes such as motion analysis and stereopsis are probably not strictly independent modules of the visual system. It is likely that different modules interact with one another. There are, however, great advantages in treating the human visual system as a set of (relatively) independent functional modules at the computational level, the output of one forming the input to others. This is only an approximation to the true organization of the system, but served as a convenient starting point for our research. For this reason, we have also concentrated our efforts on the earlier stages in the visual system rather than to proceeding too quickly to the later stages.

Though incomplete, the study of early vision is also the closest to yielding an analysis at all levels, from the theory of the information processing tasks, to the level of physiology and anatomy. The intermediate level of the algorithms used by the visual system is crucial, and often the most difficult to analyze, because the algorithms are constrained both by the computation being performed *and* the available hardware. As a consequence, psychophysical and neurophysiological data have an important role in studying the computations underlying the first steps of vision. It will help in understanding this point if I outline one such attempt in which I was involved, where the search for a neural algorithm started from the information processing level, namely the extraction of contours from the image, and their use in stereopsis, using diverse psychophysical and physiological pieces of evidence.

What is Stereopsis?

Stereopsis, which refers to the use of two slightly different viewpoints of the world to compute depth, is one of the main processes responsible for two cameras deciphering the 2-D images in terms of the 3-D surfaces that are at their origin. Stereopsis has been chosen as a framework for this account of early visual processing for a number of reasons. Not the least of these is the role it has played in the work on vision at MIT. in particular, it has stimulated a close investigation of how the retinal intensity array may be optimally represented for subsequent processing, which will be a main focus of my description. Stereopsis is also in itself an important module in the process of seeing, and one which is deceptively simple at first glance. As with so many other visual tasks that humans perform easily and effortlessly, the development of automatic systems of stereoscopic vision, which would yield immediate and important applications has proven surprisingly difficult. Finally, stereopsis seems a good choice because there exists a large body of critical psychophysical evidence, with which to define and constrain the problem.

Stereopsis arises from the fact that our two eyes view the same scene from slightly different angles. One can easily experience directly this binocular disparity by looking at objects not too distant and noting their different relative positions when closing each eye in turn. The eyes converge slightly so that the two axes of vision meet at one of the points. This point is said to be fixated by the eyes and projects to the center of vision, i.e., the center of the fovea, of each retina. Any neighboring point in the visual field will project to a point on each retina some distance away from the center of vision, but the distance will not in general be the same for each eye. In fact, the distance disparity will vary with the depth of the point in the visual field relative to the fixated point, and also with its angular distances from it. The human visual system is capable of using this disparity to recover the 3-D structure of visible surfaces from their 2-D projections.

At this point the problem of stereopsis might appear to be a relatively straightforward matter of trigonometry. We might be tempted to write a computer program to solve it, or to look inside the optic ganglia of the brain for neurons that respond preferentially to, say, a bar of light on a dark background at a certain distance. However, no such computer program for stereopsis has yet been constructed that performs at a level comparable with human vision, and although such neurons can indeed be found (see the article by Pettigrew, Scientific American, Aug '72), they have not helped us to understand (or to program on a computer) stereopsis. Our own facility to perform stereopsis has led us to gloss over what is the central difficulty of the task, as we may now see if we formally set out the main steps involved. There are four:

(1) A location in space must be selected from one retinal image.

(2) The same location in space must be identified in the other retinal image.

(3) The positions in the two images of the two corresponding points must be measured.

(4) From their relative positions in the two images, the distance of the point must be calculated.

The last two steps recover the distance to a point in the scene from the positions of its projection on the two retinas. This is a problem involving the geometry of the imaging situation that we know well how to solve. As C. Longuet-Higgins in Sussex, J. Mayhew in

Sheffield and others have shown, the problem can be solved even when we know very little about the position of the eyes or the cameras. For simplicity, I will consider a somewhat ideal situation in this article: when the observer fixates a distant point, corresponding locations lie on horizontal lines at the same vertical position in the two images. This situation is common in most psychophysical experiments. In aerial photogrammetry it is common to *rectify* the two images in order to satisfy this condition. (As we will see later, this also simplifies the second of the above four steps.) The problem of calculating distance is then very easy: an estimate can be computed by our brain or a computer just from the positions of the points in the two images in terms of simple trigonometry.

In the first two steps shown above, identifying a "location in space" means, in effect, a point on a surface. Each photoreceptor in the retina can be thought of as looking along a line of sight to a point on the surface of some object. Geometrically corresponding photoreceptors in the two retinae will not, in general, be looking at the same piece of surface. How then are corresponding pieces of surface to be identified? This, is where the difficulty of stereopsis lies. For us, the visual environment contains surfaces that are effectively "labeled" point by point by their relationships to distinct objects and shapes. But, the influential work of B. Julesz tells us at the outset a very important fact about stereopsis, namely that it does not necessarily depend upon the prior perception of objects. Julesz's work with the random dot stereogram is familiar to many Scientific American readers (see Sci. Amer. Feb., 1965). Each eye is presented with an array of dots whose patterns are almost identical but which conceal a simple 3D shape. The patterns are individually meaningless to the eye – the shape is revealed only after the two images have been fused by the brain. This finding spoke strongly against the prevailing theories of the time that gave object and shape recognition a key role in visual processing. Most importantly, Julesz' random dot stereograms allows one to formulate one computational goal of human stereopsis as the extraction of disparity information from a pair of images, without the need of obvious monocular cues.

The main problem that human stereo vision must solve is therefore what has been called the *correspondence problem* – how to find corresponding points in the two images *without* recognizing objects or their parts. This is also the main problem that has hampered the development of completely automatic stereo systems, despite many attempts. In random dot stereograms, black dots in one image are all the same, of the same size and contrast; any given dot in one image could in practice be matched with any one of a large number of dots in the other image. And yet our brain solves the false target problem and comes up with the right answer. How does the brain determine what corresponds to what?

The problem could in principle be solved in two ways. First, some other form of local labeling could be used. Each point in an image is the projection along a line of sight of a point on a physical surface (unless the camera is looking at the sky!). If a point is visible in both eyes or cameras, then its projections are the corresponding points in the two images should be matched. One candidate for points to be matched would be raw light intensity, but computer experiments show limitations to the effectiveness of this, and psychophysical evidence speaks against it for the human visual system. Indeed, points on specular surfaces will not necessarily reflect the same light intensity to both eyes, and, more important, widely separated points may have the same intensity, thus causing ambiguity. Rather than using the raw intensity array, stereopsis may operate on an intermediate representation of the image, that makes explicit primitive features that correspond more closely to physical features in space. It is this possibility that I shall now examine, beginning with some observations about the physical world.

Constraints on the Stereo Computation

The task of identifying corresponding locations in the two images is difficult because of the so-called false target problem. This seems particularly obvious in random dot stereograms in which each of the black points in one eye's view could match any of the points in the other eye's view. There are many thousands of possible matchings between the left and right images, yet our brain consistently chooses only one.

For our visual system to find a single answer to the correspondence problem, it must make use of implicit assumptions about the physical world and how it is encoded onto the eyes that allow it to determine one solution most appropriate from a physical point of view. These assumptions must be powerful enough to yield constraints that make the problem determined and solvable. There are two broad types of constraints on the stereo computation; those on the matching process itself, and those on the nature of the items - of the measurements on the image - that have to be put in correspondence. I will first discuss briefly these two types of constraints and then show how they can be used to answer the two basic question posed by the correspondence problem: *what* to match and *how* to match.

In 1976, D. Marr and I found that simple properties of physical surfaces could constrain the correspondence problem sufficiently for the stereo algorithms that we were then investigating. These are: (1) that a given point on a physical surface has only one 3D location at any given time, and (2) that objects are cohesive (and opaque) and that therefore the variation in depths over a surface are generally smooth, with discontinuous changes occurring only at surface boundaries. The uniqueness constraint means that each item in either image has a unique disparity and can be identified with no more than one location in the other image. The second, continuity, enables us to group points in space together (exploiting some of their geometry) and thus bring configurational features of the two stereo images to the aid of stereopsis. These two simple constraints provide matching rules that are reasonable and powerful, and as we will see later, lead to the correct correspondence. More refined matching rules can be formulated by considering in more detail the geometry of stereopsis. For example, if the fixation point is sufficiently distant, corresponding points have to be sought only on horizontal lines at the same height. If the surface is continuous and is not transparent, the uniqueness and continuity constraints imply that points on a horizontal line in the left image must be matched to points in the right image occurring in the same order - from left to right. As shown by T. Binford at Stanford (with H. Baker and Arnold) and J. Mayhew (with J. Frisby) in Sheffield, these and other matching rules can be usefully incorporated into the computation of stereopsis.

It is necessary first, however, to identify items for the matching process that are in one-to-one correspondence with well-defined locations on a physical surface, are stable as much as possible against photometric and geometric distortions (i.e., they appear the same in the two eyes despite the slightly different point of view), and are specific enough to simplify the matching problem. We have already seen that raw intensity values themselves are too unreliable to be used for matching. A class of items that are still quite simple, yet correspond more closely to locations on physical surfaces are surface markings, and edges. These markings and edges are of course encrypted in the grey level array provided by the sensors; to use them as matching items, they must be decoded by appropriate operations. But how? If an additional constraint on the nature of physical surfaces is added, the problem becomes simpler. This is based on the observation that at locations where there are physical changes in a surface, the image usually shows sharp changes in intensity. Intensity changes due to markings on surfaces - such as texture edges - are sufficiently prevalent in the real world to be a potential aid to stereopsis. Thus instead of the raw numerical values corresponding to the intensity values in the image, we want the stereo matching process to operate on a more symbolic, compact and robust representation. Attractive primitive symbols for this description - in a sense its basic "alphabet" - are the intensity changes. D. Marr at MIT coined the term "primal sketch" to describe a representation of different measurements of intensity changes in the original grey level array.

I will next present a scheme for detecting and describing intensity changes that we have been using at the MIT Artificial Intelligence Laboratory for the past several years, based on old and new ideas developed by several people. It has many attractive features: it works well, it is simple and it shows interesting analogies with biological vision from which it was in fact suggested. It is not, however, the full solution to this problem. From a biological point of view it is at present only a working hypothesis, certainly incomplete, at least in the simplified form that I am going to describe.

Locating Edges in Images

Changes of in an image can be detected by comparing neighboring values in the array; if the difference is large, the intensity changes very rapidly. This operation can be regarded as taking the first derivative of the image, and looking at its extremal values - either peaks or troughs. The position of the peak (or the through) localizes the position of an intensity change, which in turn often corresponds to an edge on a physical surface. A brief thought shows that an edge also corresponds to a zero-crossing in the second derivative of the image (which can be obtained by taking differences between neighboring values in the first derivative), that is the location where the second derivative crosses zero, going from positive to negative values or vice versa. Thus, through their extrema and zeros, derivatives of an image seem to provide a good means for detecting edges. Unfortunately, this scheme is still too simple to work on real images, largely because intensity changes in an image are rarely like the clean and sharp step change from one intensity value to another shown in the figure. First, there are many types of changes, slow and sharp ones, taking place over different characteristic *spatial scales*. Second, changes in intensity are also often corrupted by *noise* that infiltrates at different stages during the process of transducing the image formed by the optics of the eye or camera into an array of measurements of light intensity. Modern digital cameras as well as our eye are subject to this, often unavoidable, noise.

To cope with intensity changes over different scales and added noise, the image must be smoothed by local averaging of neighboring intensity values. The averaging procedure removes minor intensity changes due to noise, and depending on the spatial extent of the averaging, it can emphasize intensity changes at different scales. The differencing operation can then be performed on the smoothed image. There are various ways of performing these two steps and much theoretical work has been done to characterize optimal ways of doing it. (Recent work by J. Canny and by V. Torre and myself has clarified further the exact form of the optimal operations for edge detection.) An example, one of the simplest, is the scheme that we have used at M.I.T.. It was suggested to D. Marr and myself in 1977, while we were working on the problem of stereopsis, by a combination of psychophysical and physiological data on primate vision, and was successively modified and extended in an influential paper by D. Marr and E. Hildreth.

In this scheme the two operations - smoothing and differentiation - are combined into a single operation of filtering (in this case convolving) the image with a circularly symmetric spatial filter, technically called a point-spread function, whose shape is the Laplacian of Gaussian. The Laplacian-of-a-Gaussian filter, shaped like a Mexican hat, is similar to the center-surround receptive fields of retinal cells, well known to visual physiologists. Convolving an image with a filter is equivalent to substituting each value of the image with a weighted addition of neighboring values, where the weights are provided by the point-spread (or filter) function. The point-spread function shows how a single point of light would spread out in the filtered array. Once this filtering operation is performed for each image element, the result is an array of positive and negative numbers, a kind of second derivative of the image intensity. Zero-crossings in the filtered array correspond to points in the image where intensity is changing most rapidly. A binary map of the filtered array in which the positive and negative regions are represented as white and black, is completely equivalent to the map of the zero-crossings.

In the human visual system, most of the elements required to make this hypothesis workable seem to be present. As early as 1865, E. Mach observed that our perceptual system seems to enhance the changes in light intensity, and postulated the existence of a mechanism of lateral inhibition for performing an operation similar to a spatial derivative. Furthermore, there is evidence suggesting that the primate retina may do something like center-surround filtering. The output from each eye is conveyed to the brain by about a million nerve fibers, which are the axons of the retinal ganglion cells. In the retina the intensity values measured by the photoreceptors are processed by several types of neurons before they arrive at the ganglion cells. Physiological data suggest that a subclass of retinal ganglion cells, whose receptive fields have a center-surround organization (that is, an excitatory center and an inhibitory surround), closely approximate the Laplacian-of-a-Gaussian filter. Thus the array of one type of ganglion cells may represent the result of filtering the image through the center-surround filter. It is not too unreasonable to think that positive values in the filtered image would be carried by the ON-center cells and negative values by the OFF-center cells. (ON-center cells are stimulated by a bright spot in the center of the receptive field, whereas OFF-center cells are excited by a dark spot.) This is now a conjecture that connects the theory of the information processing task performed at this early stage of vision with the biological hardware that implements this task. It offers, of course, many exciting possibilities. One which I cannot refrain from mentioning is based on the recent finding by Nelson, Famiglietti and Kolb that ON- and OFF-ganglion cells (in the cat) are segregated into two different layers in the retina. The binary maps of the images provided by the MIT convolver would then literally represent activity in these two layers of cells in the retina, blue corresponding to OFF-layer activity and red to ON-layer activity.

Note that the whole operation of filtering the image is computationally very expensive for digital computers, because it involves very many multiplications: about 1 billion for an image with 1000 x 1000 elements. K. Nishihara, N. Larson and M. Kass at the Artificial Intelligence Laboratory of MIT, have designed a special hardware convolver that can perform this operation in about 1 second. This is an impressive rate but still very slow compared with the speed attained by retinal ganglion cells if they indeed perform, among other computations, the convolution of the image with a center- surround filter in real time.

Let me now briefly mention the issue of *scale*. Changes of intensity, as I mentioned, can take place over a range of spatial scales. In an image there are fine as well as coarse changes of intensity. All of these changes must be detected and represented. How can this be done? The natural solution is to use filters of different dimensions, and this is indeed what neurobiology and psychophysics (by F. Campbell and J. Robson at Cambridge University and by H. Wilson and J. Cowan of the University of Chicago) suggest to us. For a given resolution the process of finding intensity changes consists of first filtering the image with a center-surround type of filter whose extent reflects the spatial scale over which the changes must be detected, and then locating the zero-crossings in the filtered image. To detect changes at all spatial scales, it is necessary to add filters of different dimensions and carry out the same computation for each. Large filters allow the detection of soft or "blurred" edges as well as illumination changes, small ones allow the detection of fine details in the image, and all filter sizes allow the detection of high-contrast sharp edges. On grounds such as these, therefore, the hypothesis that derivative-like operations are performed on the image at different scales (or resolutions) looks attractive as a means of decoding features such as edges. The information about these edges is contained in the zero-crossings of the filtered image at each scale, which may be represented explicitly by the zero-crossing maps, or implicitly in the binary maps of the filtered image that show only the regions of positive and negative convolution values (see figures). Primitive representations of this type at different scales are a far cry from the raw intensity array and sometimes reveal hidden features in the image, as well as edges of different orientations and degrees of blur.

Recent theoretical results have enhanced the attractiveness of this idea by showing that features similar to zero-crossings in the filtered image (or to the binary maps of the filtered array) can be very rich in information. First, the mathematician B. Logan at Bell

Laboratories proved that a function filtered through a certain class of filters can be completely reconstructed from its zero-crossings alone (modulus an overall scaling factor). Though the Laplacian-of-a-Gaussian filter does not exactly satisfy Logan's conditions, the theorem suggests that the relatively sparse number of discrete "symbols" provided by the zero-crossings are very rich in information about the filtered image. In recent months A. Yuille and I have obtained specific results about the zero-crossings (and other extrema points) of images filtered with the derivatives of a Gaussian filter, such as the center-surround filter shown in the figures. Our theorems suggest that zero-crossings maps obtained at different scales, i.e. with different filter sizes, (in principle, one can think in terms of a continuum of scales, an idea first suggested by A. Witkin at the Fairchild Artificial Intelligence Laboratory) are a complete representation of the image. Clearly there is no need for reconstructing the original image from these symbols. But our theorems suggest that the features obtained by appropriately filtering the image capture a lot of information and represent, therefore, one of the candidates for an optimal encoding scheme used by later processes, such as stereopsis. It must be pointed out that the Laplacian-of-a-Gaussian filter and the associated binary convolution maps (or its zero-crossings) are just one of the representations suggested by this theoretical analysis. What exactly are the representations that have to be used by artificial vision systems and are used by our visual system is still an open question.

To summarize, the combination of computational arguments and biological data suggests that an important first step for stereopsis (and other visual processes as well) is to detect and describe the changes in intensity in the image at different scales. A representation of intensity changes can be compact and explicit, representing the information that matters for later processes. Filtering with the Laplacian-of-a-Gaussian filter is an especially simple and convenient scheme; the zeros in the filtered array essentially correspond to "edges" in the image. (Interestingly a similar operation can be useful for finding other kinds of edges, such as color boundaries. The expected color-coded center-surround cells have been recently found in special anatomic structures in the primate visual cortex by D. Hubel and M. Livingstone.) It has been known since the early work of D. Hubel and T. Wiesel that there exists cells in the primary visual cortex that respond selectively to edges of a particular orientation and sign of contrast in their receptive fields. It is unclear whether cortical cells exist that explicitly signal the presence of zero-crossing segments of a certain orientation in their receptive fields. It is unlikely that cortical cells behave as originally proposed by D. Marr, though experiments by K. Richter suggest that some cells (in the visual cortex of the cat) may indeed encode information about zero-crossings.

Note that information about the zero-crossings implicit in the pattern of activity of some of the ON- and OFF- ganglion cells layers in the retina, since the zero-crossings are the locations where activity switches from one layer to the other (we neglect in this simplified description all temporal properties of ganglion cells).To explicitly represent the zero-crossings cortical cells would be required that connect neighboring ON-center and OFF-center cells, and perform a multiplication or logical AND operation on their outputs. Variations of this scheme are also possible, including one in which the logical operation being performed is the AND-NOT operation. At this point, we feel very strongly the lack of understanding of what elementary computational operations nerve cells can readily perform. H. Barlow has suggested, in fact, that the "veto" or AND-NOT operation is one for which nerve cells may be especially adapted, and V. Torre, C. Koch and I have shown that this operation can be performed between pairs of synaptic inputs on a small patch of the dendritic membrane of a neuron. Neurons in the visual cortex are mostly likely doing much more than detecting oriented "edges" or zero-crossings or peaks in derivatives of the image. But the work of physiologists suggests that representing intensity changes in a way somewhat similar to the scheme we have discussed, may indeed be one of the tasks of some cortical cells.

Stereopsis Algorithms

In order to see how a representation of intensity changes might be used in stereopsis, we shall first consider an algorithm that was devised by D. Marr and myself in 1976, which incorporates the previously stated computational constraints, and which is successful at solving random dot stereograms (the approach was first suggested by P. Dev). As the figure shows, the algorithm requires a 3D network of nodes or "cells", each of which lies at the intersection of a line of sight from each image. Each node contains a "0" or a "1", depending on whether a correspondence is established between the two points corresponding to the two intersecting lines. To implement the uniqueness constraint, the nodes lying along a given line of sight strictly inhibit each other. To implement the continuity constraint, each node excites its immediate neighbors.

In the unnatural case of a random dot stereogram, each dot can be made to correspond to one line of sight and has one of two distinct intensities (black or white). To run the algorithm, we begin by placing a 1 in each node at which the intensities agree, and 0 at the others. Each 1 represents a match, whether the true one or a false one. In the execution of the algorithm, each node adds up the excitatory inputs from its coplanar neighbors at the same disparity, and subtracts the inhibitory inputs from its colinear ones. If the result exceeds a threshold value, the node takes the value 1, otherwise it is set to 0. After a few such steps, the network reaches stability and the problem is solved.

The cooperative algorithm has some quite considerable virtues. In particular, it is composed of local interactive operations that can run asynchronously in parallel and which could easily be identified with individual neurons. In addition, the network not only solves the false target problem, but can also "fill-in", effectively interpolating a continuous surface. Notice also that whereas the algorithm favors continuity of matches in the disparity domain (in this simple version it favors front and parallel planes) it allows for sharp discontinuities at boundaries. However, the 3D network required to process finely detailed natural images would seem to be very large, and most of the nodes in this network would be idle at any one time. Furthermore, as pointed out earlier, intensity values are not satisfactory for matching under less restricted conditions than those that apply to the random dot stereograms.

The range of effectiveness of this algorithm can be extended to more natural images by transforming the images to obtain their zero-crossings or equivalently the sign of the Laplacian-of-a-Gaussian convolution. This transformation provides us with a means for converting natural scenes into patterns that bear a resemblance to Julesz's binary random dot stereograms (the sign of this convolution array is completely equivalent to the zero-crossing map). The cooperative algorithm can now operate on this representation — this binary array — exactly as on random dot stereograms and extract the correct disparities. In addition, the convolution can be performed with different sized filters, that reveal primitive, otherwise hidden features at the different resolutions. Economies could therefore be achieved in the brain by processing the image with a range of filters of different sizes interacting in an appropriate way.

At present it is unclear whether performance of this simple version of the cooperative algorithm on natural images can be satisfactory. It is quite surprising that such a simple and "blind" algorithm relying on primitive constraints can perform so well for at least some natural images. Of course, natural images contain a variety of cues in addition to binocular disparity, like shading, contours and occlusions that human observers are very adept at using for computing depth. Julesz's demonstration that monocular information is not needed to fuse a stereogram does not imply that it is not used when available. A complete theory of human stereo vision should eventually also include ways to process this additional information.

In any case, the representation of the image in terms of the convolution with the center-surround filters at different resolutions opens up much more interesting possibilities than the use of the simple cooperative algorithm One obvious possibility is to use as matching features the zero-crossing contours themselves. At low spatial resolutions, zero-crossings of a given sign (for instance zero-crossings for which the convolution output changes from positive to negative) are quite rare and in fact never too close. Thus false targets - matches between noncorresponding zero-crossings - are essentially absent over a relatively large disparity range. A mathematical analysis of the probability of occurrence of zero-crossings in bandpass filtered images it shows that if the disparity range that is considered is on the order of the size of the receptive field used to filter the image (more precisely the diameter of the "excitatory" center), false targets are virtually absent. These observations led D. Marr and myself to propose in 1977 a different algorithm for solving the correspondence problem. In its simplest version the algorithm matches zero-crossings of the same sign in image pairs filtered with center-surround receptive fields of 3 or more different sizes. First the coarse images are matched and the disparity measured. This rough result is used to approximately register the two images in the region of interest (monocular features and texture differences can also be used). The same matching process is then applied to the medium-sized filter. A similar procedure finally yields very fine disparity resolution in the small disparity range within which the smallest, high resolution filter operates. A theoretical extension and computer implementation of this algorithm by E. Grimson at the Artificial Intelligence Laboratory (described in his book that also addresses the basic problem of surface reconstruction) shows many of the properties of human depth perception, for instance the ability to perform successfully when one of the stereo images is defocused. That there are also some significant differences is shown in recent work by J. Mayhew and J. Frisby at Sheffield and B. Julesz at Bell Laboratories.

Other stereo algorithms that make direct use of the primitive multiscale description obtained by filtering the image with the center-surround filters are also quite effective. For instance, K. Nishihara at the MIT Artificial Intelligence Laboratory has developed a stereo system for robotics applications that operates on the sign of the convolution of the image. It detects the presence of a correspondence between regions of these binary representations and measures its rough disparity in a sequence of increasingly finer resolutions. Its main characteristics are speed and capability of coping with noisy images.

At this point what can we say about the biological mechanisms of stereopsis? The algorithms that I have described are still far from solving the correspondence problem as effectively as the human visual system but suggest ways of how it may be solved at all. Experiments in the cat's visual cortex by H. Barlow, C. Blakemore and Pettigrew in 1967, and in the visual area of the macaque by D. Hubel, T. Wiesel and by G. Poggio, revealed that cortical neurons signal binocular disparity, but did not provide insight on how the brain handles the correspondence problem that occurs during binocular vision. Very recently, G. Poggio has found cells in the visual cortex of the macaque that signal the correct disparity of random dot stereograms in which there are many possible false matches. This discovery promises to give us new insights about the brain mechanisms underlying stereoscopic matching. In particular, while some disparity sensitive cells may be thought to use matching primitives that correspond to isolated oriented edges, the activity of some complex cells may reflect something similar to patchwise correlation of filtered images, an operation that is almost equivalent to matching zero-crossings over a certain area and that could be performed in the same cortical area or elsewhere.

The interaction of the information processing approach with these new physiological data is an exciting development that should raise our confidence in the prospect of understanding human stereo vision, perhaps in the near future. Work on computational algorithms for stereopsis is clarifying the central issues in terms of information processing. Psychophysics and physiology are important for solving the problem of human stereo vision, testing specific models and even for helping to develop automatic stereo systems. Many problems that I did not mention must still be solved if we want to understand human stereopsis well enough to be able to implement it in computers and robots with the same level of performance. Stereopsis is an early but difficult stage of vision; it provides a unique and exciting opportunity to demonstrate that attempts to understand the human visual system and to develop computer vision systems that can fruitfully interact.

Conclusions

One message that should have emerged from this discussion is the extent to which computers and brains can be brought together in the investigation of a problem like vision. On the one hand, computers provide a powerful tool for testing computational theories and algorithms and proving their ability (or sufficiency) to solve the problem. In the process, they serve as a guide for the design of neurophysiological experiments, suggesting what to look for. The impetus that this will give to brain research over the next decades is likely to be very great.

But the benefit is not only one way; computer science also stands to gain. It is not the case, as some computer scientists have maintained, that the brain provides no more than an existence proof, i. e. a living demonstration that a given problem has a solution. It can also, as I hope this article has suggested, show us how to seek a solution. The brain is an information processing machine that has evolved over millions of years to perform certain real-world tasks superlatively well. If we tend, with usually not unjustified modesty, to regard our brain as a somewhat uncertain instrument of reason, this is only because we are conscious of the things it does less well - the recent things in evolutionary terms like logic, mathematics and philosophy - and are normally unconscious of its true powers, like vision. It is in these functions that we have a lot to learn from the brain, and it is against these functions that we should judge our achievements in computer science and robotics. If we do this we may begin to see what vast potential lies ahead of us.

Acknowledgments: I am grateful to E. Hildreth for her very substantial help for fixing some of the bugs in this article and to H. K. Nishihara for the figures (see Scientific American) and the many discussions. Carol Bonomo should also be acknowledged for something.