themselves after migrating. The positional information in the cytoplasm may in turn depend on the products of "maternal effect" genes, maternal genes that are expressed in the egg and have among their effects the establishment of the egg's anterior-to-posterior and dorsal-to-ventral axes.

Once *ftz* and other genes that are needed to establish the basic segmentation pattern are turned on, then the homeotic genes of the Antennapedia and bithorax complexes may be activated to direct those segments to diversify along appropriate paths. Activation of the genes that finally produce the different structural features of the fruit fly would be a fairly late event, possibly the result of the activities of the homeotic gene products, according to this view.

Whether or not the homeo box functions in higher species such as the frog, mouse, and human in the way postulated for it in the fruit fly is currently unknown. However, there are indications that genes bearing the boxes may function early in development in these species. The Basel workers have shown that such a gene in the frog is transcribed into mRNA by the late gastrula stage. The presence of homeo boxes may provide a handle by which some early developmental genes can be identified and studied, which would be a big help to researchers who wish to unravel the mysteries of development.—**JEAN L. MARX**

Computer Vision

This may be as close as AI has yet come to being a true science; but even so, no one really knows what it means to "see"

Legend has it that a certain pioneer in artificial intelligence research (AI) once gave a graduate student a little project for the summer: solve vision.

That was two decades ago.

One wonders if the student had a very good time that summer. Not only is his little problem of vision still unsolved, it is still one of the greatest challenges in AI. Vision systems do exist for industrial robots, for example, yet even now they tend to be primitive silhouette matchers with limited utility. And when the Defense Advanced Research Projects Agency (DARPA) recently launched its "Strategic Computing" initiative (*Science*, 16 December 1983, p. 1213), it estimated that another 10 years of concentrated effort would be required before an autonomous reconnaissance vehicle could "see" well enough to rove over unknown terrain.

But in all fairness, the professor's overconfidence was natural. Back in the 1960's AI researchers tended to think of vision as rather easy, largely because we do it ourselves with no mental effort at all. A game like chess seemed to require much more thought, and there were already programs that could play chess passably well.

And indeed the goal of vision does seem rather straightforward. As the late David Marr of the Massachusetts Institute of Technology (MIT) recently wrote, "Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information" (*1*).

However, the simplicity is deceptive. It is one thing to record an image with a camera; it is quite another thing to understand what that image represents. In the early 1970's AI researchers began to write vision programs in earnest—and began to realize what a horrendous thing vision really is.

First, a real-world image contains an enormous amount of data, much of it irrelevant and all of it subject to noise and distortion. In practice this means that a vision system has to have huge amounts of memory and processing power. If one begins with a high-resolution image measuring 1000 by 1000 pixels—a "pixel" being a single digitized picture element—even some of the simplest procedures require about 100 million operations. The human retina, which has approximately 100 million rods and cones, plus four other layers of neurons, all operating at roughly 100 hertz, performs at least 10 billion calculations per second before the image even gets to the optic nerve. And then, once the image information reaches the brain, the cerebral cortex has more than a dozen separate vision centers to process it. In fact, from studies on monkey brains it has been estimated that vision in one form or another involves some 60 percent of the cortex.

The upshot is that if seeing seems effortless, it is because we do not have to think about it; the whole massive computation is unconscious. If chess seems hard, it is only because we *do* have to think about it.

Second, one has the ironic fact that with all this information, there is still not enough. An image is just the two-dimensional projection of a three-dimensional world; the reverse transformation, from the 2-D image to the 3-D objects, is highly ambiguous. So far as a 2-D image on the retina is concerned, for example, the family cat might as well be carved into the tip of an infinitely long rod directed straight away from the eye. And yet, because we know that cats are not like that we never perceive the poor beast that way. Clearly, a competent vision system needs to "know" about cats, and dogs, and an enormous variety of other things, just to resolve the ambiguities.

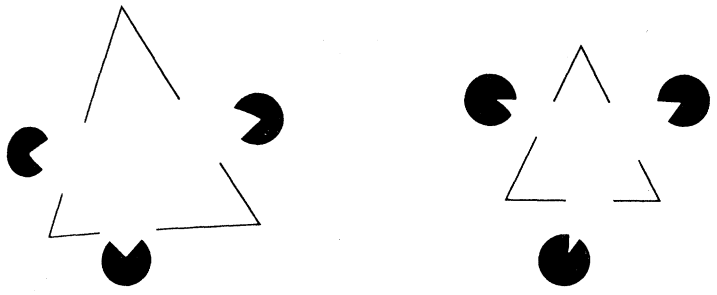
Third, an object may only vaguely resemble others of its generic type. Consider a real cat, a porcelain cat, and a cat made out of twisted pipe cleaners: What is it that allows us to recognize them all as cats? In addition, as lighting conditions or viewing angles change, an object may not even resemble itself; consider a cat as seen from the side, and a cat as seen face on. This fact alone makes the commercial "template-matching" vision systems hopelessly inadequate for anything but the carefully controlled environment of a factory.

Finally, there are a myriad of possible objects in the world, and almost as many generic types. Humans can handle them all, in principle. A powerful vision system should be able to do it too.

Laid out like this, the problem of vision might seem hopeless. But, in fact, the computer vision community is surprisingly optimistic. The next few years promise to bring an enormous increase in computational power, largely due to the development of a new class of processors that do their calculations in parallel instead of in series.

But perhaps more important, there is a sense in the community that the "low-level," or "early" part of the vision problem, the perception of 3-D shape from 2-D imagery, is well on its way to a systematic solution.

The most influential single figure in this development was MIT's David Marr, whose career was cut short in 1980 by his death from leukemia at age 35. Not everyone was fond of Marr: by all reports he was highly articulate, deliberately provocative, and something of a showman. But his synthesis of computational AI and the experimental psychology of vision, leavened with the mathematical work of such researchers as MIT's Berthold Horn, was undeniably a landmark in the field. Marr's nomenclature for early vision has become ubiquitous. And even the people who disagree with his theories often feel obliged to refer to them as a point of contrast. As Marr's MIT colleague Tomaso Poggio puts it, "One of David's main achievements was to convince people in AI that there is a lot of *science* to be done in early vision—as opposed to a lot of ad hoc hacking."



Subjective contours

Objectively this is nothing but an arrangement of lines and blobs. But the visual system seems to think otherwise. Surfaces and changes in depth are so important in the world that we have evolved a neural algorithm to fill in missing contours; thus it is almost impossible to see these figures as anything but white, curvilinear shapes masking complete balls and triangles below. [Courtesy of W. H. Freeman & Company]

Marr gave a highly readable summary of his views in his book *Vision* (*1*), published posthumously in 1982. His basic idea was that before a vision system can jump to identifying objects, it must first identify surfaces with definite positions and orientations in space. He drew a sharp distinction between this approach and the "segmentation" algorithms that were popular in the mid-1970's. The idea there was to have the programs segment the image into blobs of near equal shading or intensity and then try to identify the original objects from an analysis of the blobs. Segmentation never worked too well, said Marr, because the most dramatic boundaries in an image often come from such irrelevant factors as reflections or shadows, while many of the most important edges may be nearly invisible. By first identifying surfaces, he argued, a program could avoid such illusions.

Drawing on a large body of psychophysical evidence, Marr then argued that in human vision the identification of surfaces begins very early, sometimes before the visual information even leaves the retina for the brain. The neurons of the retina and the visual cortex appear to embody a number of quasi-independent algorithms—"modules"—that exploit such clues as texture, color, motion, shading, or, in the case of stereo vision, parallax. In the process, these modules attempt to resolve the two-dimensional ambiguities of the image by making certain general assumptions about the world. For example, the "motion" module appears to assume that surfaces are rigid; thus, if a swarm of details moves uniformly across our field of view, we perceive those details as lying on a surface. The "stereo" module assumes that surfaces are continuous and opaque, and so on.

Being hardwired into the system, so to speak, these assumptions stamp themselves on our perceptions remorselessly—even when the assumptions are wrong. Thus we have optical illusions. But such occasional mistakes are the price of breaking out of two dimensions, argued Marr. Constraints are essential and should be a prime subject of study in computer vision.

Finally, he said, the key problem in early vision is to integrate the output of these modules into a coherent picture of the surfaces and their relationships to the viewer. To see how nontrivial this is, consider a photograph: our stereo vision tells us it is flat, and yet we ignore that and heed the cues of shape and shading that tell us, "This is a person's face."

The integration begins at an intermediate level of processing that Marr called the "2½-D" sketch. This is the high end of immediate perception, in that there is little information left to be extracted from the image alone. In subsequent processing the visual system begins to incorporate higher order world knowledge to form a full 3-D representation of objects—to say, for example, that such and such a collection of surfaces is in fact a cat instead of a Christmas tree.

Marr's preference for bringing in knowledge at this level, rather late in the processing, was in contrast to the then current tendency of computer vision researchers to incorporate world knowledge from the very beginning—the "I see a cat because I *expect* to see a cat" approach. A number of researchers, such as Horn, were already arguing that this approach violates both the psychophysical evidence and common sense. Why make a commitment to what is out there until you have extracted as much information from the image as possible? So Marr called this "the principle of least commitment," and made it one of the foundations of his theory.

Marr's program thus called for three distinct stages of processing: two-dimensional, "2½-" dimensional, and three-dimensional. His illness cut short his work on the latter two. But for those first, earliest phases of vision, he and his colleagues at MIT were able to turn their theoretical ideas into a remarkable series of working algorithms (*2*).

The first step in their approach was to transform the raw data from the input image—one intensity value for each pixel—into a more compact and symbolic representation in terms of edges and intensity variations. Marr called this representation "the primal sketch."

Among other things, getting to the primal sketch means filtering out noise from the image and simultaneously enhancing broad-scale changes in intensity. There are innumerable ways to do this, of course, but the MIT group chose to concentrate on a particular filter function known as the Laplacian of a Gaussian, which has a graph that resembles a Mexican hat. There were good biological reasons for doing so: such a function seems to correspond to the receptive fields of certain retinal ganglion cells, in which brightness in the center of the field excites the cell while brightness just off-center inhibits it. In any case, the result is a filtered image that appears to be almost nothing but edges, often in startling relief.

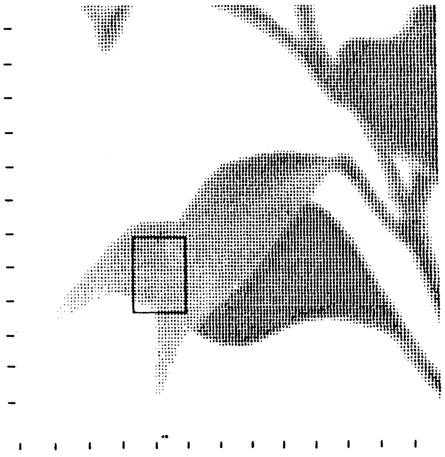
Once the edges are found they can be used as input for many of the constraint algorithms—"modules" such as stereopsis, or the determination of structure from motion. In the case of stereopsis, for example, a program might use edges to make a first rough match between points in the left image and equivalent points in the right image.

The upshot of all this work is that a sizable piece of the early vision problem is now in reasonably good shape, especially when it comes to identifying the assumptions that make a specific problem solvable, and implementing those assumptions in working algorithms. In addition to MIT there are currently at least a dozen university groups doing research on vision, plus about half-a-dozen industrial laboratories; the list of working "modules" includes not only stereopsis (uniqueness and continuity of surfaces) and structure from motion (rigidity), but such things as the filling in of missing contours (edges tend to be smooth and continuous), surface interpolation (surfaces tend to be smooth and continuous), and the location of surfaces from texture (surface elements such as polka dots tend to be uniformly distributed and of equal size).

Moreover, in an interesting recent development, Poggio and some of his coworkers have begun to explore the so-called "regularization" techniques of mathematical physics; they may offer a way of putting all, or almost all, of the early vision problems into a common framework. The idea is to formulate the constraints assumed by the vision system as a certain kind of problem in the calculus of variations, which may then suggest more effective algorithms for solving the problem. "In the last few years, there has been a lot of work in bringing mathematical and physical techniques to bear," says Poggio. "Computer vision is becoming much more technical as it becomes closer to being a science—to showing us how we see and how we can make a computer see; that's why it's so exciting."

Unfortunately, as even Marr had to admit, vision as a whole is a long way from being solved. Even at the level of the 2-D sketch, for example, it is not obvious what kind of "surface" information the eye is extracting from a transparent windowpane or from the reflections on a rippling pool of water. And what does one do with a bank of fog or a wisp of smoke—things that do not even have surfaces?

Matters only get worse as one proceeds into the higher stages of visual processing, the 2½-D and 3-D sketches. There the programs must begin to incorporate real-world knowledge, which means that the programmers have to start grappling with all the old imponderables of knowledge representation and computer cognition (*Science*, 23 March, p. 1279). Such high-level activities are far less accessible to experiment than early vision, and thus there is far less guidance on how to proceed.



More subjective contours

We have no trouble perceiving this as an image of two leaves—but where, exactly, is the boundary between them? Inside the box, at least, there really is no boundary, which means that our visual systems are doing a lot more than just analyzing blobs of light and dark. [Courtesy of W. H. Freeman & Company]

As an example, consider the ubiquitous problem of representation. Before a program can recognize a three-dimensional object, or reason about it, that program clearly has to have something in memory to compare it with—a kind of Platonic archetype, perhaps.

But consider what is involved in our own perception of, say, a cat. Whatever our internal representation may be, it is certainly specific enough to distinguish a cat from a dog or a goat. But it is also general enough to encompass a standing cat, a running cat, a sleeping cat, a Picasso cat, a pipe-cleaner cat, and all the other reasonable variations. The representation clearly cannot depend upon the cat's having fur and vertically slit eyes because otherwise what would we make of a porcelain cat? And yet when we see a living cat we never perceive its gray-striped coat as something separate. By the time perceptions reach our conscious minds, the general and the specific have long since been integrated.

The problem is that knowing all this tells us nothing about how to achieve the same kind of versatility in a computer—especially in a computer with finite memory capacity. People have tried a number of approaches, notably the "generalized cone" system pioneered by Thomas Binford of Stanford University. (Objects are approximated as a collection of cones and cylinders with the appropriate shapes, sizes, and orientations; a stick figure just traces out the axes of the cones.) But the fact is that no one has yet come up with a satisfactory representation scheme.

"It is fairly clear that the human mind has multiple representations for everything," says John Kender of Columbia University. "That has good survival value. But it's my belief that it doesn't just reduce down to [the equivalent of] symbolic LISP statements. It may possibly be some kind of 'analogic' representation, in which a 3-D object corresponds to a complex interconnection of neurons in the brain."

Problems like these are clearly part of the long haul for computer vision research. As Kender says, "Even if we had machines a million times faster than we have now, we still wouldn't know how to write the algorithms." Genuine progress in high-level vision is probably going to take at least another decade.

On the other hand, it is also true that new forces are at work that promise to speed up the pace considerably. Most notable are DARPA's Strategic Computing initiative, and the advent of very fast, massively parallel computers.

DARPA, of course, has been AI's major funding source from the beginning, which has led many in the AI community to see Strategic Computing as a mixed blessing. DARPA's reassurances notwithstanding, there is, for example, some concern that the new initiative will cut into AI's basic research funds, especially as the agency tries to live up to its ambitious promises. Already, skeptics claim to see a suddenly increased emphasis on real-time program performance as opposed to a more fundamental understanding of intelligence—an emphasis that could ultimately be self-defeating.

And yet, the DARPA initiative promises to bring new money and new energy to the development of parallel processors, with computer vision as one of the major beneficiaries. This in itself will not guarantee any breakthroughs in fundamental understanding, of course. But it will certainly help. A number of such systems will become operational over the next 2 years or so, and the computer vision community is deeply involved in planning ways to use them.

—M. Mitchell Waldrop

This is the fourth in a series of occasional articles on artificial intelligence. Previous articles appeared in Science, *24 February, p. 802; 23 March, p. 1279; and 27 April, p. 372. A subsequent article will deal with advanced computer architectures and parallel processing.*

Additional Reading

1. D. Marr, *Vision* (Freeman, San Francisco, 1982).
2. Poggio gives a popular account of some of this work in *Scientific American* **250**, 106 (April 1984).