Vision

1992 McGraw-Hill Yearbook of Science and Technology

Donald D. Hoffman

Department of Cognitive Science

University of California, Irvine, 92717

From the viewpoint of artificial intelligence, a vision system is a sophisticated computing system with one primary purpose: to construct useful descriptions of the environment from images. These descriptions typically focus on objects and their interrelationships—the three-dimensional shapes of objects, their motions relative to each other and to the viewer, the manner in which their surfaces absorb and scatter light, their arrangement in space, their textures, and their identities. The problems solved by a vision system, as it constructs these descriptions, are the reverse of those solved by a computer graphics system. Whereas a graphics system starts with a model of a three-dimensional world which it then must render as a two-dimensional image or sequence of images, a vision system, by contrast, starts with the two-dimensional images and must recover descriptions of the three-dimensional world and its objects.

Images. The inputs to a vision system are images, either a discrete sequence of images much like the discrete frames of a videotape, or a continuously varying family of images. In the discrete case, an image is a two-dimensional array of integers, say $I(x, y)$. If the image is black and white, then each integer represents a shade of gray at a unique point in the image. This is perhaps most easily understood by looking closely at a picture from a newspaper. Inspection reveals that the picture is actually composed of a two-dimensional array of dots, varying from very dark (larger dots) to very light (smaller dots). The array of dots is integrated by the eye and perceived as a coherent image. The case of color images follows the same principle, but is a bit more complex. Instead of one integer for each point in the image there are now three, indicating the intensities of red, green, and blue at each point. A common example of this idea is the image on a color television. In either case, for images that are in color or in black and white, the problem to be solved by a vision system is plainly quite difficult: given a two-dimensional array of numbers (one row of which might be, for example: 231, 235, 223, 229, . . .) determine the corresponding state of the three-dimensional environment (for example, recognize that the image is a picture of a child riding a bicycle).

Edges. Once a vision system has acquired an image, perhaps by means of a video camera, it has several tasks to perform. One task is to locate and to classify edges and line segments in the image. Of the many approaches to this problem developed recently, perhaps the best known proceeds by convolving (filtering) the image with a two-dimensional gaussian $G(x, y)$, and then taking the laplacian $\nabla^2 = \partial^2/\partial^2 x + \partial^2/\partial^2 y$ of the result. These two steps can be combined into one operation, namely convolution of the image with the function $\nabla^2 G$ (pronounced "del square g"). This function has the shape, roughly, of a Mexican hat. Once the image has been convolved with a $\nabla^2 G$, the result can then be used to find edges and line segments: these correspond roughly to locations where the numerical values in the convolved image pass through zero, so-called “zero crossings”. This process is illustrated in Figure 1. By changing the variance (roughly, the width or dispersion) of the gaussian used in the $\nabla^2 G$ one can look for edges at different scales of resolution: higher variance for coarser resolution, and lower variance for finer resolution. Once a vision system has obtained the edges in an image, it then must interpret them, for example by classifying them as arising from shading, object boundaries, surface markings (such as the grain of wood), or texture boundaries. This classification problem has as yet not been solved in general, though progress has been made on special cases.

Three-dimensional structure. To facilitate navigation through the environment, and to allow the manipulation and recognition of objects in that environment, a vision system must infer the shapes, locations, and motions of objects in three dimensions. How this inference can be performed, quickly and reliably, has been the subject of much recent research. Among the “cues to depth” in images that have received attention are visual motion, stereo, shading, texture, surface contours, and occluding contours (roughly, the silhouettes of objects). In each case researchers have found that the information in two-dimensional images, by itself, is insufficient to dictate a unique and reliable three-dimensional interpretation. In consequence, a vision system must bring to bear background constraints or assumptions to guide the interpretation process. This has been a key insight.

Consider, for instance, the “cosine surfaces” depicted in Figure 2. The pattern of curves shown in the figure is certainly planar, since it is printed on the page. However, it is nearly impossible to avoid perceiving (or, more precisely, *misperceiving*) the curves as lying on two nonplanar surfaces in three dimensions. This illustrates that human vision readily infers three dimensions from two, and that its inferential processes can at times reach incorrect conclusions. What is perhaps remarkable here is not simply that human viewers misperceive the figure, but that they all report the *same* misperception. This suggests that, of the many three-dimensional interpretations one could give to the figure, human vision consistently employs some assumption or constraint to pick out a unique interpretation. It has been proposed, for instance, that human vision interprets this figure by assuming that the curves lie on a surface for which the curves are in fact lines of curvature (in the differential geometric sense). This proposal may or may not be correct in detail, but it illustrates a strategy—namely the search for reasonable assumptions or constraints on the set of possible interpretations—that has repeatedly led to progress in the study of machine and human vision.

Usually the constraints employed by human vision lead to a small number of consistent interpretations. Occasionally they fail to do so, with amusing results as can be seen in Figure 3.

Motion. One powerful cue to depth is visual motion. From a sequence of images, such as the successive frames of a videotape, in which objects move slightly from frame to frame, it is often possible to perceive not simply two-dimensional motions, but in fact three-dimensional motions and three-dimensional shapes. This ability in human vision is so well developed that it typically permits a person to drive a car safely even if one eye is not functioning. Recent theoretical work on this topic, usually called the perception of “structure from motion”, has resulted in several powerful theorems which state conditions in which one can obtain (1) a unique three-dimensional interpretation and (2) a low probability of false interpretations. Here is one such theorem: If a vision system is given three distinct images of four or more feature points, points which are moving about in three dimensions, then (1) if the images allow the points to be interpreted as moving on a *rigid* object in three dimensions then, generically, they allow at most two such interpretations, and (2) the probability is zero that the imaged points can be interpreted as moving on a rigid object in three dimensions, given that in fact they were not so moving. This theorem, and others of a similar nature, provide a theoretical foundation upon which to build reliable machine vision systems for the interpretation of visual motion. These theorems also permit the principled development of noise insensitive algorithms for the recovery of three-dimensional structure from image motion.

Manipulation and recognition. Motion, stereo, shading, and other cues allow a vision system to infer the three-dimensional structure of objects. But how this three-dimensional structure should be represented, once it has been inferred, depends critically upon the tasks for which it is to be used. If, for example, one wishes to use the three-dimensional information to grasp and manipulate the object, then it is preferable to use a "viewer-centered" representation, one which makes explicit the distance and orientation of the object relative to the viewer, for it is the three-dimensional shape and position of the object relative to the viewer that is critical for this task. If, on the other hand, one wishes to recognize the object, then it is preferable to use an "object-centered" representation, one which makes explicit the intrinsic geometry of the object independent of its position relative to the viewer, so that small changes in viewing position do not fundamentally change the description of the object and thereby impede the recognition process. Researchers in machine vision continue to refine both kinds of representation, and have successfully built systems, for certain restricted environments, that permit manipulation and recognition. The problem of devising general purpose vision systems for recognition and manipulation is much more difficult and not yet solved.

Prospects. Recent advances in our understanding of vision are due in large part to increased multidisciplinary collaborations between experts in visual neurobiology, artificial intelligence, mathematics, and visual psychophysics. Continued research along these lines promises both technological and scientific advances. Technologically, the increasing speed of digital hardware, coupled with the new theories and algorithms being developed by vision researchers, are making possible a more versatile generation of machine vision systems. As these systems increase in sophistication they are likely to see broader application: in automated assembly, in the exploration of space, as the eyes of home and industrial robots, and as prosthetic devices for the visually impaired. Scientifically, we can expect continued progress in constructing mathematically rigorous theories and theorems (such as the theorem described in the section on motion) for various aspects of visual information processing—from the detection of edges to the recovery and recognition of three-dimensional shapes. Moreover we can expect to see general principles emerge, principles which when properly formalized and investigated may well transform the scientific study of vision.

Bibliography.

1. D. Marr, *Vision,* 1982.
2. D. Ballard and C. Brown, *Computer Vision,* 1982.
3. B. Horn, *Robot Vision,* 1985.
4. J. Wolfe, *The Mind's Eye,* 1986.
5. W. Richards, *Natural Computation,* 1988.
6. B. Bennett, D. Hoffman, and C. Prakash, *Observer Mechanics,* 1989.

7. J. Aloimonos and D. Shulman, *Integration of Visual Modules,* 1989.

Figure Legends

Figure 1. Finding edges in an image. Part **a** shows the $\nabla^2 G$ function in profile; the full function is obtained by simply rotating this figure a half turn about a vertical line through its middle. Part **b** shows a section through an image containing an edge; where the curve is higher the image is brighter, and where the curve is lower the image is darker. The transition from light to dark is an edge. Part **c** shows the result of convolving **a** with **b** (the convolution being indicated by *). The point where the curve in **c** passes through the horizontal line is the zero crossing that indicates the existence and location of the edge.

Figure 2. Two planar figures that appear three-dimensional. Observe how the hills and valleys exchange places when the figure is turned upside down.

Figure 3. The "devil's tuning fork". The assumptions used by human vision to analyze this figure lead to no globally consistent interpretation in three dimensions.