Computer Vision 1

Computer Vision Issues

1.1 ACHIEVING SIMPLE VISION GOALS

Suppose that you are given an aerial photo such as that of Fig. 1.1a and asked to locate ships in it. You may never have seen a naval vessel in an aerial photograph before, but you will have no trouble predicting generally how ships will appear. You might reason that you will find no ships inland, and so turn your attention to ocean areas. You might be momentarily distracted by the glare on the water, but realizing that it comes from reflected sunlight, you perceive the ocean as continuous and flat. Ships on the open ocean stand out easily (if you have seen ships from the air, you know to look for their wakes). Near the shore the image is more confusing, but you know that ships close to shore are either moored or docked. If you have a map (Fig. 1.1b), it can help locate the docks (Fig. 1.1c); in a low-quality photograph it can help you identify the shoreline. Thus it might be a good investment of your time to establish the correspondence between the map and the image. A search parallel to the shore in the dock areas reveals several ships (Fig. 1.1d).

Again, suppose that you are presented with a set of computer-aided tomographic (CAT) scans showing "slices" of the human abdomen (Fig. 1.2a). These images are products of high technology, and give us views not normally available even with x-rays. Your job is to reconstruct from these cross sections the three-dimensional shape of the kidneys. This job may well seem harder than finding ships. You first need to know what to look for (Fig. 1.2b), where to find it in CAT scans, and how it looks in such scans. You need to be able to "stack up" the scans mentally and form an internal model of the shape of the kidney as revealed by its slices (Fig. 1.2c and 1.2d).

This book is about *computer vision*. These two example tasks are typical com-

puter vision tasks; both were solved by computers using the sorts of knowledge and techniques alluded to in the descriptive paragraphs. Computer vision is the enterprise of automating and integrating a wide range of processes and representations used for vision perception. It includes as parts many techniques that are useful by themselves, such as *image processing* (transforming, encoding, and transmitting images) and *statistical pattern classification* (statistical decision theory applied to general patterns, visual or otherwise). More importantly for us, it includes techniques for geometric modeling and cognitive processing.

1.2 HIGH-LEVEL AND LOW-LEVEL CAPABILITIES

The examples of Section 1.1 illustrate vision that uses *cognitive processes*, *geometric models*, *goals*, and *plans*. These *high-level* processes are very important; our examples only weakly illustrate their power and scope. There surely would be some overall purpose to finding ships; there might be collateral information that there were submarines, barges, or small craft in the harbor, and so forth. CAT scans would be used with several diagnostic goals in mind and an associated medical history available. Goals and knowledge are high-level capabilities that can guide visual activities, and a visual system should be able to take advantage of them.

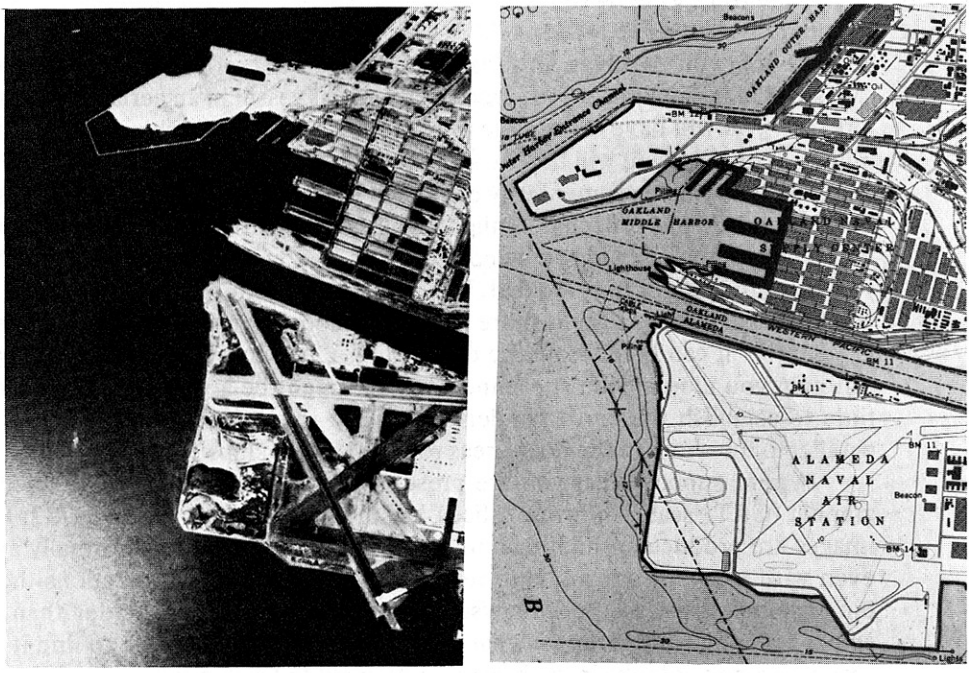
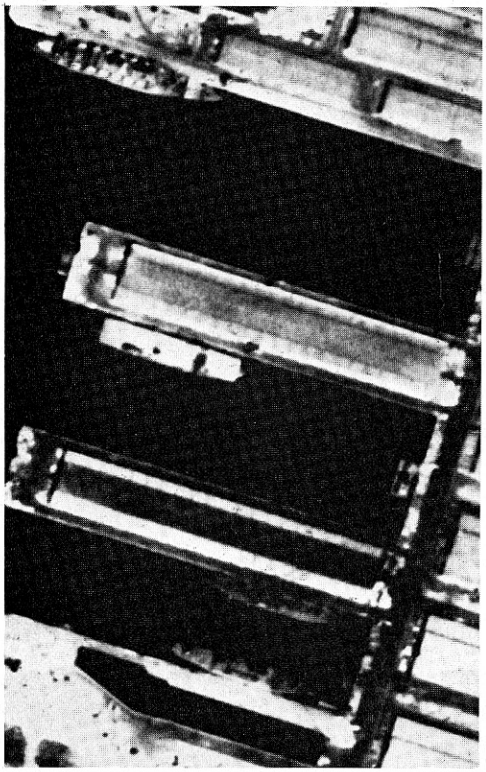
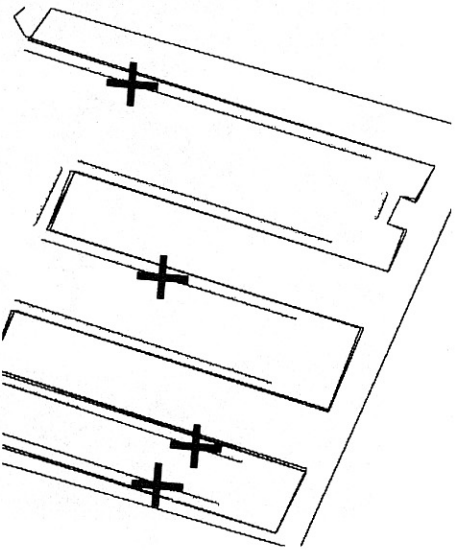


Fig. 1.1 Finding ships in an aerial photograph. (a) The photograph; (b) a corresponding map; (c) the dock area of the photograph; (d) registered map and image, with ship location.



(c)



(d)

Fig. 1.1 (cont.)

Even such elaborated tasks are very special ones and in their way easier to think about than the commonplace visual perceptions needed to pick up a baby, cross a busy street, or arrive at a party and quickly "see" who you know, your host's taste in decor, and how long the festivities have been going on. All these tasks require judgment and large amounts of knowledge of objects in the world, how they look, and how they behave. Such high-level powers are so well integrated into "vision" as to be effectively inseparable.

Knowledge and goals are only part of the vision story. Vision requires many *low-level* capabilities we often take for granted; for example, our ability to extract *intrinsic images* of "lightness," "color," and "range." We perceive black as black in a complex scene even when the lighting is such that some black patches are reflecting more light than some white patches. Similarly, perceived colors are not related simply to the wavelengths of reflected light; if they were, we would consciously see colors changing with illumination. Stereo fusion (stereopsis) is a low-level facility basic to short-range three-dimensional perception.

An important low-level capability is *object perception*: for our purposes it does not really matter if this talent is innate, ("hard-wired"), or if it is developmental or even learned ("compiled-in"). The fact remains that mature biological vision systems are specialized and tuned to deal with the relevant objects in their environ-

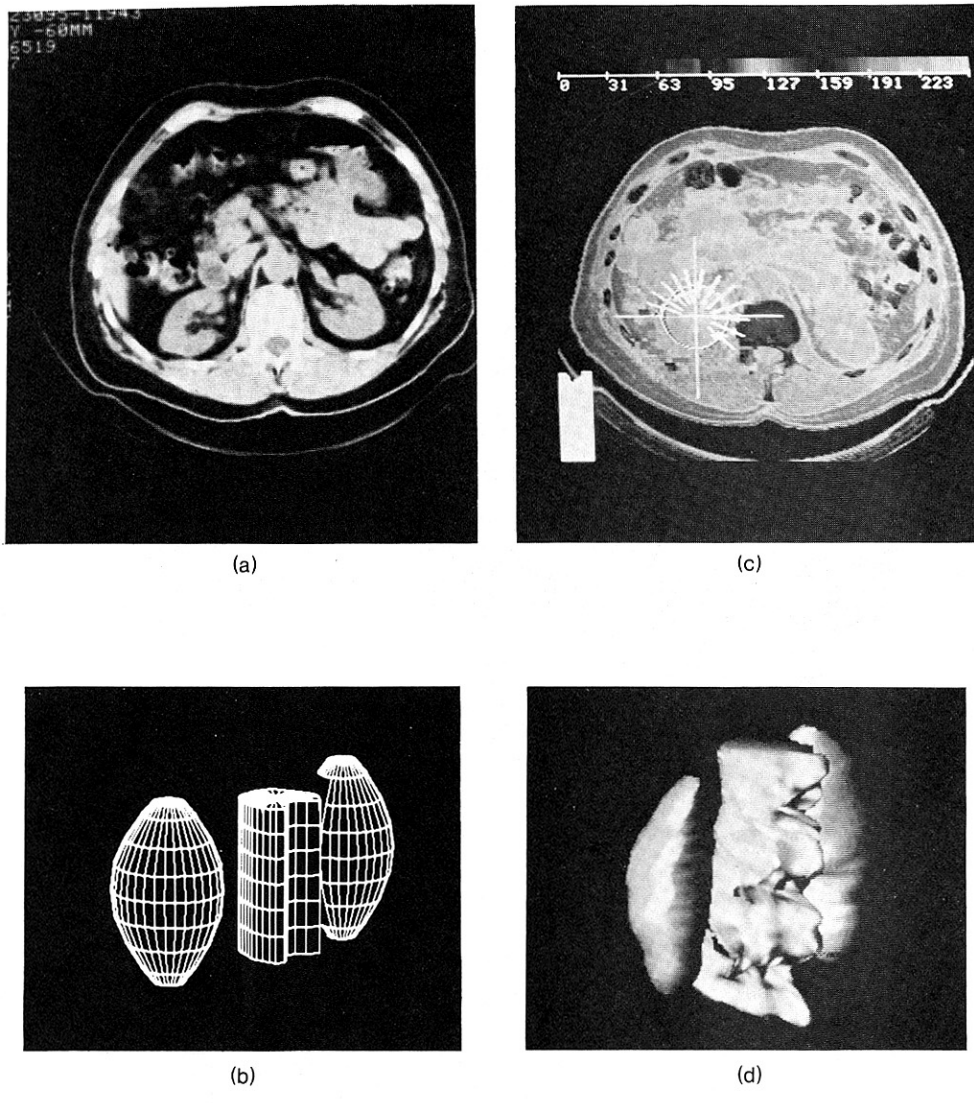


Fig. 1.2 Finding a kidney in a computer-aided tomographic scan. (a) One slice of scan data; (b) prototype kidney model; (c) model fitting; (d) resulting kidney and spinal cord instances.

ments. Further specialization can often be learned, but it is built on basic immutable assumptions about the world which underlie the vision system.

A basic sort of object recognition capability is the "figure/ground" discrimination that separates objects from the "background." Other basic organizational predispositions are revealed by the "Gestalt laws" of clustering, which demonstrate rules our vision systems use to form simple arrays of stimuli into more coherent spatial groups. A dramatic example of specialized object perception for

human beings is revealed in our "face recognition" capability, which seems to occupy a large volume of brain matter. Geometric visual illusions are more surprising symptoms of nonintuitive processing that is performed by our vision systems, either for some direct purpose or as a side effect of its specialized architecture. Some other illusions clearly reflect the intervention of high-level knowledge. For instance, the familiar "Necker cube reversal" is grounded in our three-dimensional models for cubes.

Low-level processing capabilities are elusive; they are unconscious, and they are not well connected to other systems that allow direct introspection. For instance, our visual memory for images is quite impressive, yet our quantitative verbal descriptions of images are relatively primitive. The biological visual "hardware" has been developed, honed, and specialized over a very long period. However, its organization and functionality is not well understood except at extreme levels of detail and generality—the behavior of small sets of cat or monkey cortical cells and the behavior of human beings in psychophysical experiments.

Computer vision is thus immediately faced with a very difficult problem; it must reinvent, with general digital hardware, the most basic and yet inaccessible talents of specialized, parallel, and partly analog biological visual systems. Figure 1.3 may give a feeling for the problem; it shows two visual renditions of a familiar subject. The inset is a normal image, the rest is a plot of the intensities (gray levels) in the image against the image coordinates. In other words, it displays information

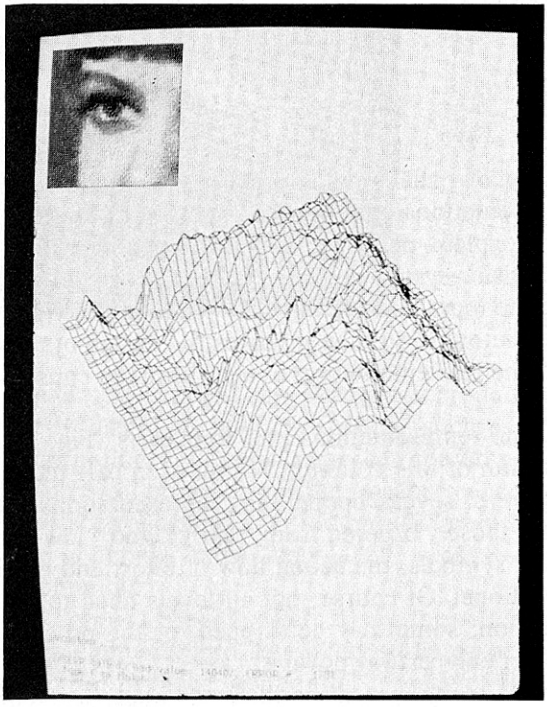


Fig. 1.3 Two representations of an image. One is directly accessible to our low-level processes; the other is not.

with "height" instead of "light." No information is lost, and the display is an image-like object, but we do not immediately see a face in it. The initial representation the computer has to work with is no better; it is typically just an array of numbers from which human beings could extract visual information only very painfully. Skipping the low-level processing we take for granted turns normally effortless perception into a very difficult puzzle.

Computer vision is vitally concerned with both low-level or "early processing" issues and with the high-level and "cognitive" use of knowledge. Where does vision leave off and reasoning and motivation begin? We do not know precisely, but we firmly believe (and hope to show) that powerful, cooperating, rich representations of the world are needed for any advanced vision system. Without them, no system can derive relevant and invariant information from input that is beset with ever-changing lighting and viewpoint, unimportant shape differences, noise, and other large but irrelevant variations. These representations can remove some computational load by predicting or assuming structure for the visual world.

Finally, if a system is to be successful in a variety of tasks, it needs some "meta-level" capabilities: it must be able to model and reason about its own goals and capabilities, and the success of its approaches. These complex and related models must be manipulated by cognitive-like techniques, even though introspectively the perceptual process does not always "feel" to us like cognition.

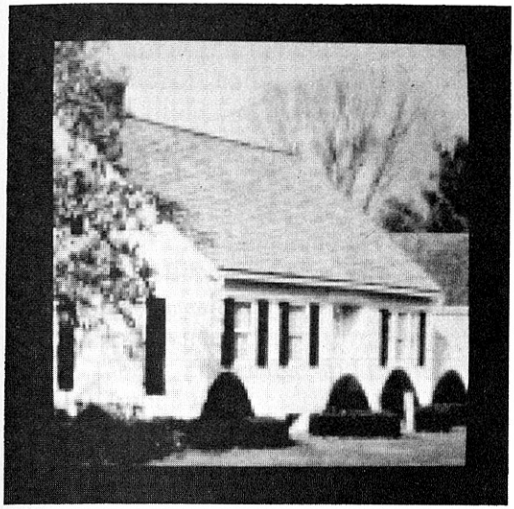
Computer Vision Systems

1.3 A RANGE OF REPRESENTATIONS

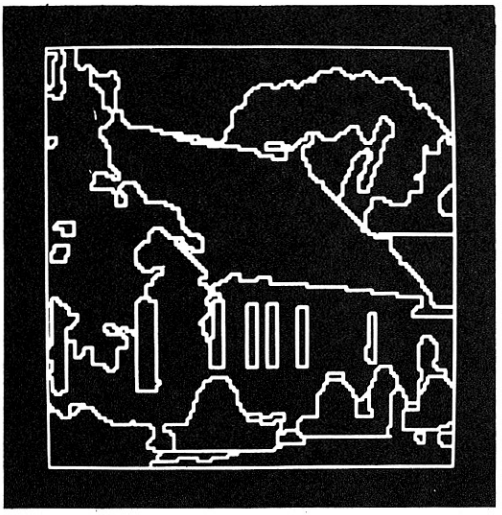
Visual perception is the relation of visual input to previously existing *models* of the world. There is a large representational gap between the image and the models ("ideas," "concepts") which explain, describe, or abstract the image information. To bridge that gap, computer vision systems usually have a (loosely ordered) *range of representations* connecting the input and the "output" (a final description, decision, or interpretation). Computer vision then involves the design of these intermediate representations and the implementation of algorithms to construct them and relate them to one another.

We broadly categorize the representations into four parts (Fig. 1.4) which correspond with the organization of this volume. Within each part there may be several layers of representation, or several cooperating representations. Although the sets of representations are loosely ordered from "early" and "low-level" *signals* to "late" and "*cognitive*" symbols, the actual flow of effort and information between them is not unidirectional. Of course, not all levels need to be used in each computer vision application; some may be skipped, or the processing may start partway up the hierarchy or end partway down it.

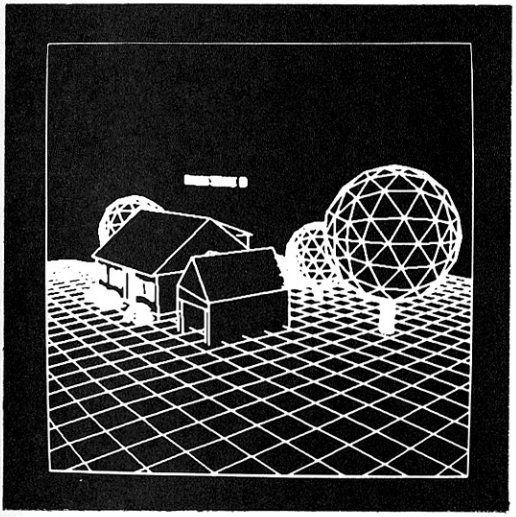
Generalized images (Part I) are *iconic* (image-like) and *analogical* representations of the input data. Images may initially arise from several technologies.



(a)



(b)

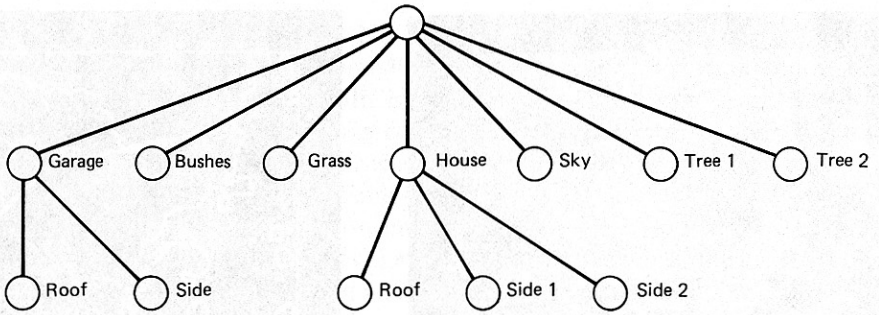


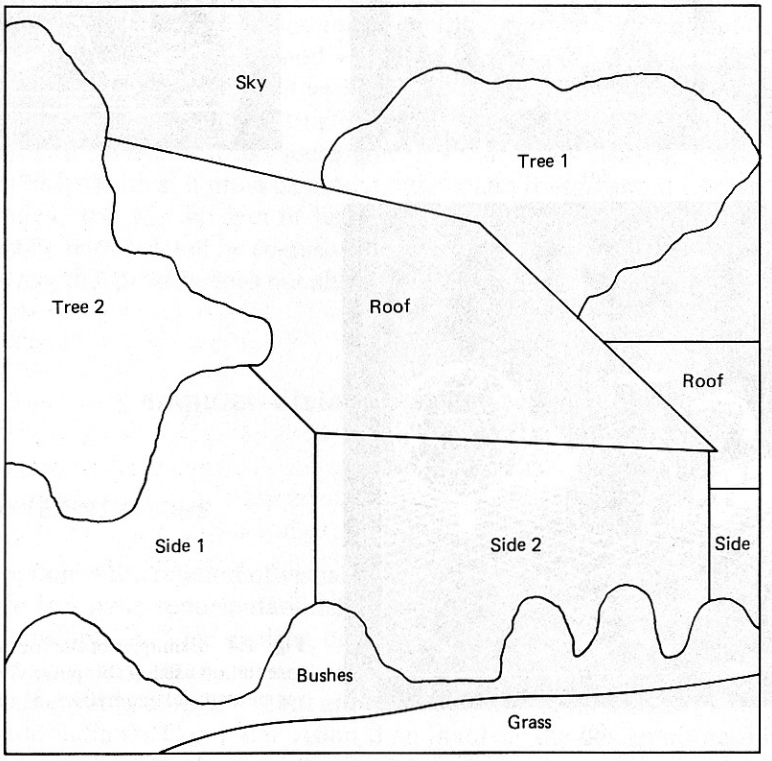
(c)

Fig. 1.4 Examples of the four categories of representation used in computer vision. (a) Iconic; (b) segmented; (c) geometric; (d) relational.

Domain-independent processing can produce other iconic representations more directly useful to later processing, such as arrays of *edge elements* (gray-level discontinuities). *Intrinsic images* can sometimes be produced at this level—they reveal physical properties of the imaged scene (such as surface orientations, range, or surface reflectance). Often *parallel processing* can produce generalized images. More generally, most "low-level" processes can be implemented with parallel computation.

Segmented images (Part II) are formed from the generalized image by gathering its elements into sets likely to be associated with meaningful *objects* in the scene. For instance, segmenting a scene of planar polyhedra (blocks) might result in a set of *edge segments* corresponding to polyhedral edges, or a set of two-





(d)

Fig. 1.4 (cont.)

dimensional *regions* in the image corresponding to polyhedral faces. In producing the segmented image, knowledge about the particular domain at issue begins to be important both to save computation and to overcome problems of *noise* and inadequate data. In the planar polyhedral example, it helps to know beforehand that the line segments must be straight. *Texture* and *motion* are known to be very important in segmentation, and are currently topics of active research; knowledge in these areas is developing very fast.

Geometric representations (Part III) are used to capture the all-important idea

of two-dimensional and three-dimensional *shape*. Quantifying shape is as important as it is difficult. These geometric representations must be powerful enough to support complex and general processing, such as "simulation" of the effects of lighting and motion. Geometric structures are as useful for encoding previously acquired knowledge as they are for re-representing current visual input. Computer vision requires some basic mathematics; Appendix 1 has a brief selection of useful techniques.

Relational models (Part IV) are complex assemblages of representations used to support sophisticated high-level processing. An important tool in *knowledge representation* is *semantic nets*, which can be used simply as an organizational convenience or as a formalism in their own right. High-level processing often uses prior knowledge and models acquired prior to a perceptual experience. The basic mode of processing turns from *constructing* representations to *matching* them. At high levels, *propositional* representations become more important. They are made up of assertions that are true or false with respect to a model, and are manipulated by rules of *inference*. Inference-like techniques can also be used for *planning*, which models situations and actions through time, and thus must reason about temporally varying and hypothetical worlds. The higher the level of representation, the more marked is the flow of *control* (direction of attention, allocation of effort) downward to lower levels, and the greater the tendency of algorithms to exhibit *serial processing*. These issues of control are basic to complex information processing in general and computer vision in particular; Appendix 2 outlines some specific control mechanisms.

Figure 1.5 illustrates the loose classification of the four categories into analogical and propositional representations. We consider generalized and segmented images as well as geometric structures to be analogical models. Analogical models capture directly the relevant characteristics of the represented objects, and are manipulated and interrogated by simulation-like processes. Relational models are generally a mix of analogical and propositional representations. We develop this distinction in more detail in Chapter 10.

1.4 THE ROLE OF COMPUTERS

The computer is a congenial tool for research into visual perception.

- Computers are versatile and forgiving experimental subjects. They are easily and ethically reconfigurable, not messy, and their workings can be scrutinized in the finest detail.
- Computers are demanding critics. Imprecision, vagueness, and oversights are not tolerated in the computer implementation of a theory.
- Computers offer new metaphors for perceptual psychology (also neurology, linguistics, and philosophy). Processes and entities from computer science provide powerful and influential conceptual tools for thinking about perception and cognition.
- Computers can give precise measurements of the amount of processing they

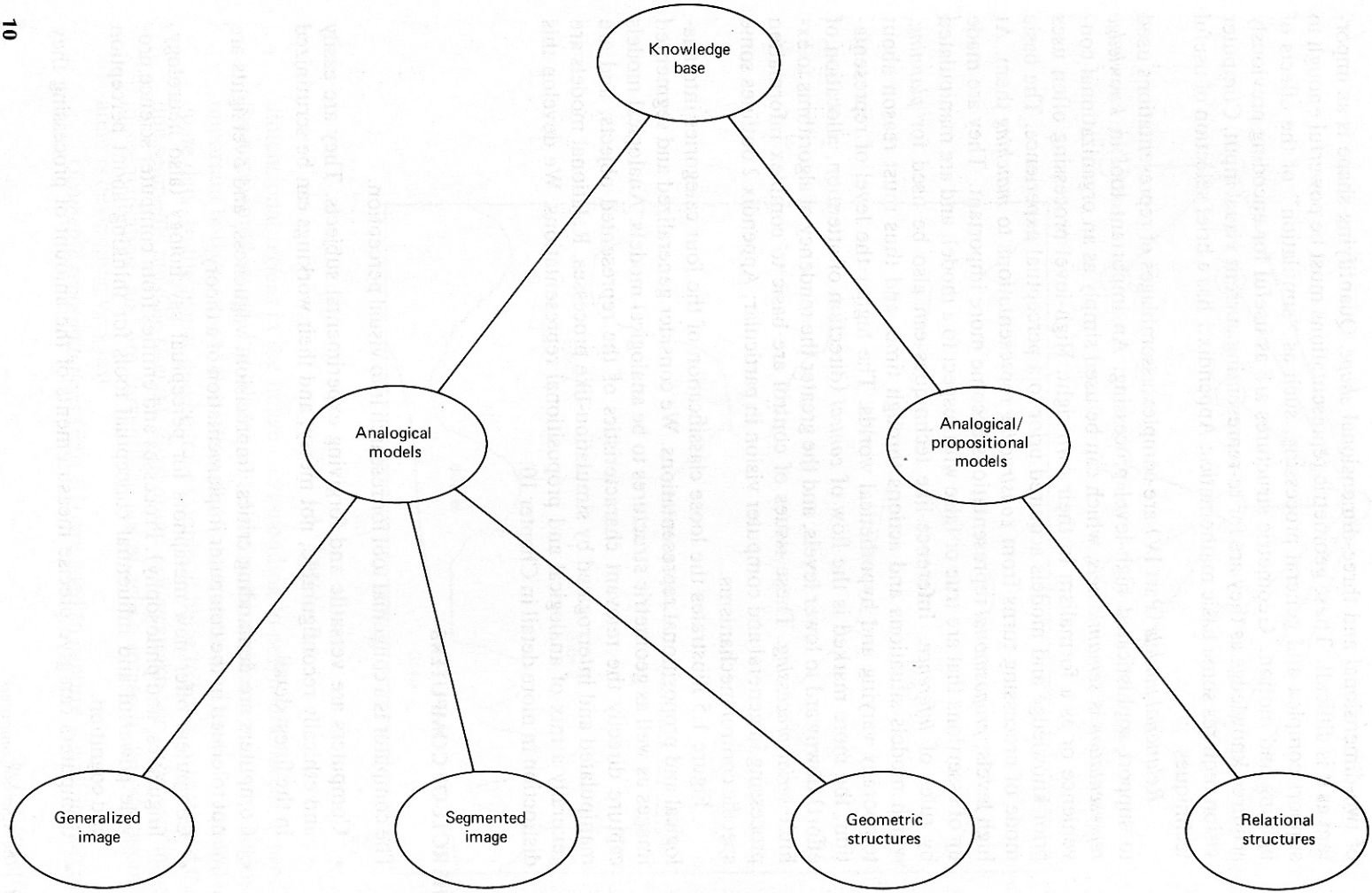


Fig. 1.5 The knowledge base of a complex computer vision system, showing four basic representational categories.

Table 1.1

EXAMPLES OF IMAGE ANALYSIS TASKS

Domain	*Objects*	*Modality*	*Tasks*	*Knowledge Sources*
Robotics	Three-dimensional outdoor scenes indoor scenes	Light X-rays	Identify or describe objects in scene Industrial tasks	Models of objects Models of the reflection of light from objects
	Mechanical parts	Light Structured light		
Aerial images	Terrain Buildings, etc.	Light Infrared Radar	Improved images Resource analyses Weather prediction Spying Missile guidance Tactical analysis	Maps Geometrical models of shapes Models of image formation
Astronomy	Stars Planets	Light	Chemical composition Improved images	Geometrical models of shapes
Medical Macro	Body organs	X-rays Ultrasound Isotopes Heat	Diagnosis of abnormalities Operative and treatment planning	Anatomical models Models of image formation
	Cells	Electronmicroscopy	Pathology, cytology	
Micro	Protein chains Chromosomes	Light	Karyotyping	Models of shape
Chemistry	Molecules	Electron densities	Analysis of molecular compositions	Chemical models Structured models
Neuroanatomy	Neurons	Light Electronmicroscopy	Determination of spatial orientation	Neural connectivity
Physics	Particle tracks	Light	Find new particles Identify tracks	Atomic physics

do. A computer implementation places an upper limit on the amount of computation necessary for a task.

- Computers may be used either to mimic what we understand about human perceptual architecture and processes, or to strike out in different directions to try to achieve similar ends by different means.
- Computer models may be judged either by their efficacy for applications and on-the-job performance or by their internal organization, processes, and structures—the theory they embody.

1.5 COMPUTER VISION RESEARCH AND APPLICATIONS

"Pure" computer vision research often deals with relatively domain-independent considerations. The results are useful in a broad range of contexts. Almost always such work is demonstrated in one or more applications areas, and more often than not an initial application problem motivates consideration of the general problem. Applications of computer vision are exciting, and their number is growing as computer vision becomes better understood. Table 1.1 gives a partial list of "classical" and current applications areas.

Within the organization outlined above, this book presents many specific ideas and techniques with general applicability. It is meant to provide enough basic knowledge and tools to support attacks on both applications and research topics.