ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE SECOND EDITION

VOLUME 1

Stuart C. Shapiro, *Editor-in-chief*

A Wiley-Interscience Publication
John Wiley & Sons, Inc.
New York / Chichester / Brisbane / Toronto / Singapore

IMAGE UNDERSTANDING

Think about the process by which you understand what you see. Can you determine what is happening and how it is happening when you look out the window and notice that your best friend is walking toward your door? As you may guess, the process by which you arrived at this conclusion, and which caused you to go and open the door before your friend knocked, is not a simple one. Ancient philosophers worried about this problem. Biological scientists have been studying the problem in earnest since Hermann von Helmholtz (1821–1894), commonly credited as the father of modern perceptual science. Computer scientists began looking at this problem only recently in these terms, and the discipline of computer vision is a very young one. The miracle of vision is not restricted to the eye; it also involves the cortex and brain stem and requires interactions with many other specific brain areas. In this sense, vision may be considered an important aspect of AI. It is the major source of input for man's other cognitive faculties.

This article discusses the aspects of vision that deal with the understanding of visual information. Understanding in this context means the transformation of visual images (the input to the retina) into descriptions of the world that can interface with other thought processes and elicit appropriate action. The representation of these descriptions and the process of their transformation are not currently understood by the biological sciences. In AI, researchers are concerned with the discovery of computational models that behave in the same ways that humans do, and thus, representations and processes are defined using the available computational tools.

Image understanding (IU) is the research area concerned with the design and experimentation of computer systems that integrate explicit models of a visual problem domain with one or more methods for extracting features from images and one or more methods for matching features with domain models using a control structure. Given a goal, or a reason for looking at a particular scene, these systems produce descriptions of both the images and the world scenes that the images represent.

The goal of an image-understanding system (IUS) is to transform two-dimensional spatial (and, if appropriate to the problem domain, time-varying) data into a description of the three-dimensional spatio-temporal world. The description can take many forms, and the particular form associated with a given implementation depends strongly on the problem domain and task involved; it may range from simple "yes-no" answers to full surface reconstructions of objects and anything in between. In the early to mid-seventies, this activity was termed *scene analysis*. Other terms for this are *knowledge-based vision* or *high-level vision*. IU is distinguished from *model-based vision*, whose main goal is to locate specific models and derive their transformation parameters in images (see OBJECT RECOGNITION). The descriptions sought in IUSs are more general. Several survey papers have appeared on this topic. The interested reader is particularly referred to papers by Binford (1982), Kanade (1977), Matsuyama (1984), and Tsotsos (1984) as well as the excellent collection of papers in *Computer Vision Systems* (Hansen and Risemon, 1978) and *Readings in Computer Vision* (Fischler and Firschein, 1987). Those readers interested in the biological side of image understanding are referred to an excellent book by Uttal (1981), *A Taxonomy of Visual Processes*.

Integration is the key phrase when describing an IUS. Research on IUSs has experimented with ways of integrating existing techniques into systems and, in doing so, has discovered problems and solutions that would not otherwise have been uncovered. Integrated within a single framework, an IUS must:

Extract Meaningful Two-Dimensional (2–D) Grouping of Intensity-Location-Time Values. Images or image sequences contain a tremendous amount of information in their raw form. The process of transformation thus begins with the identification of groups of image entities, pixels. These pixels are grouped by means of similarity of intensity value, for example, over a particular spatial location. They can also be grouped on the basis of intensity discontinuity or similarity of change or constancy over time. The assumption is that groups of pixels that exhibit some similarity in their characteristics probably belong to specific objects or events. Typical groupings are edges, regions, and flow vectors.

Infer 3–D Surfaces, Volumes, Boundaries, Shadows, Occlusion, Depth, Color, Motion. Using the groupings of pixels and their characteristics, the next major transformational step is to infer larger groupings that correspond, for example, to surfaces of objects or motion events. The result of the inference may be quantitative or qualitative depending on the problem task. The reason for the need for inference is that the pixels by themselves do not contain sufficient information for the unique determination of the events or objects; other constraints or knowledge must be applied. This knowledge can be of a variety of forms, ranging from knowledge of the imaging process including viewpoint, knowledge of the image formation process including camera or sensor motion, and knowledge of physical constraints on the world, to knowledge of specific objects being viewed. Typically, the most appropriate knowledge to use is an open question, but the simplest and least application-specific knowledge is preferred.

Group Information Into Unique Physical Entities. Surfaces can be connected to form 3–D objects, and changes in trajectories can be joined to describe motions of specific types. Again, the original pixel values do not contain sufficient information for this process, and additional knowledge must be applied. This knowledge is perhaps in the form of connectivity and continuity constraints, and in many cases these are embedded in explicit models of objects of the domain.

Transform Image-Centered Representations Into World-Centered Representations. To this point the descriptions created have all been in terms of a coordinate system that is "image centered" (also called "viewer centered" or "retinotopic"). A key transformation is to convert this coordinate system to one that is "world centered" (also called "object centered"), that is, the description is no longer dependent on specific locations in images. This is a crucial step—otherwise, the stored models must be replicated for each possible location and orientation in space.

Label Entities Depending on System Goals and World Models. It almost never occurs that humans are given a picture or told to look out the window and asked to describe everything that is seen in a high and uniform degree of detail. Typically a scene is viewed for a reason. What exactly this goal is has direct impact on how the scene is described, which objects and events are described in detail; and which are not. Second, scenes are always described based on what is known about the world; they are described in terms of the domain that is being viewed. A factory scene, for example, is almost never described in terms of a hospital environment—that would not be a useful description (unless metaphoric use is the goal!). This knowledge base permits the choice of the most appropriate "labels" to associate with objects and events of the scene. Labels are typically the natural-language words or phrases that are used in the applications domain. The process of finding labels and their associated models that are relevant is called "search." Models that are deemed relevant may be termed "hypotheses." Each hypothesis must be "matched" against the data extracted from the images. In the case where the data is insufficient to verify a model, "expectations" may be generated that guide further analysis of the images. Labels are necessary for communication to other components of a complete intelligent system that must use interpreted visual information. The label set forms the language of communication between vision and the remainder of the intelligent system.

Infer Relationships Among Entities. In viewing a scene, not only are individual objects and events recognized but they are also interrelated. Looking out the window, for example, one may see a tree in a lawn, a car on a driveway, a boy walking along the street, or a girl playing on a swing set. The relationships may play an important role in assisting the labeling process as well. These relationships form a spatio-temporal context for objects and events.

Construct a Consistent Internal Description. This really applies to all levels of the transformation process that is being described here. The output of an image-understanding system is a representation of the image contents, usually called an "interpretation." Care is required, however, in defining what an interpretation actually involves. Little attention has been given to this, and current systems employ whatever representation for an interpretation is convenient and appropriate to the problem domain. Basically, an interpretation consists of inferred facts, relationships among facts, and representations of physical form. Issues of consistency and foundations of the underlying representational formalism are important, yet they have not received much attention with the IUS community. The output of an IUS usually takes one of two forms: a graphic rendition of the objects recognized is displayed, perhaps with natural-language labels identifying various parts, or textual output describing the characteristics of the objects observed and recognized is generated. Some systems employ both methods, and the choice depends on the particular problem domain being addressed.

Two basic questions arise when describing an IUS to the uninitiated. The first question is "Why did this field arise as distinct from so-called low-level vision or early vision?" There are two main reasons for the distinction: the bottom-up approach (see PROCESSING, BOTTOM-UP AND TOP-DOWN) embodied in early vision schemes is inadequate for the generation of complete symbolic descriptions of visual input, and there is a need to describe visual input using the same terminology as the problem domain. There are several basic realities that impact the design of image-understanding systems. The first is that images underconstrain the scenes that they represent. The reason is straightforward: in human vision, a 3-D scene undergoes a perspective projection onto a 2-D retina in order to become an image. Thus, much information is lost, particularly depth information. The image is just a snapshot in time of the scene, and both spatial as well as temporal continuity information is lost. Further, the image created is a distorted view of the scene that it represents. The distortion is not only due to the perspective transformation, but, also, there is noise involved in the image creation process. Finally, a purely bottom-up (or data-directed) approach does not lead to unambiguous results in all cases. A data-directed scheme considers all the data and tries to follow through on every hypothesis generated. Consideration of all data and all possible models in a system of size and scope comparable to the human visual system leads to combinatorial explosion and is thus an intractable approach. Moreover, it can be nonconvergent, can only produce conclusions that are derivable directly or indirectly from the input data, and cannot focus or direct the search toward a desired solution.

A vision system must be able to represent and use a very large number of object and event models. If the input is naturally ambiguous, a purely bottom-up activation of models will lead to a much larger set of models to consider than is necessary or salient. The working hypothesis of IUSs is that domain knowledge (qv), in addition to the bottom-up processes, can assist in the disambiguation process as well as reduce the combinatorial problem. How that knowledge is to be used is a key problem.

The second question that often arises is "Is image understanding computationally the same as speech understanding?" On the surface, it may seem that the techniques applicable to the speech understanding (qv) problem are directly applicable to the image-understanding problem. A simplified view of the speech understanding process leads to this conclusion. The differences arise if content is considered, rather than form alone. Speech understanding (qv) may be regarded as the recognition of phonemes, the grouping of phonemes into words, the grouping of words into sequences, the parsing of word sequences into sentences, and the interpretation of the meaning of the sentences. Indeed, in a paper by Woods (1978), the similarity is presented in some detail. Woods speculates on the applicability of the HWIM architecture for the image-understanding problem and concludes that it may be worth the attempt. However, a closer examination of the differences between speech and image interpretation tasks reveals that the image-understanding task is significantly different and more difficult.

The similarities between the speech and image tasks are many. Both domains exhibit inherent ambiguity in the signal, and thus signal characteristics alone are insufficient for interpretation. Reliability of interpretation can be increased by the use of redundancy provided by knowledge of vocabulary, syntax, semantics, and pragmatic considerations; and both domains seem to involve a hierarchical abstraction mechanism. The differences include the facts that: (a) speech exhibits a single spatial dimension (amplitude) with a necessary temporal dimension, whereas images display two spatial dimensions as well as the temporal dimension; (b) a speech segment has two boundary points, whereas an image segment, as a spatial region, has a large number of boundary points; (c) speech has a relatively small vocabulary that is well documented (eg, in dictionaries) and images have much larger, undocumented vocabularies; (d) grammars have been devised for languages, but no such grammars exist for visual data; (e) although speech differs depending on the speaker, images vary much more because of viewpoint, illumination, spatial position, and orientation of objects, and occlusion; (f) speech has a convenient and well-accepted abstract description, namely, letters and words, whereas images do not; and (g) the speech signal is spatially one-dimensional, and when sampled by the ear, there is no equivalent of the projection of a 3-D scene onto a 2-D retina. Thus, it seems that the image-understanding situation is radically different, particularly in combinatorial terms, and it is for this reason that very different solutions have appeared.

REPRESENTATIONAL AND CONTROL REQUIREMENTS

This section attempts to summarize the experience of the IU community in the design and implementation of IUSs with a statement of components currently believed to be necessary for vision systems. It should be clear that this is not a formal definition of an IUS in a strict sense; many of the requirements are really topics for further research. The section does not contain specific references; instead, it refers to other entries in this encyclopedia. Specific solutions and vision systems and how they deal with each of these requirements appear in a subsequent section.

Representational Requirements

Many IUSs distinguish three levels of representations: a low level, an intermediate level, and a high level. These levels do not necessarily refer to particular types of formalisms but rather simply point out that in the interpretation process, a transformation of representations into more abstract ones is required and that typically three levels of abstraction are considered. These levels can usually be characterized as follows: Low level includes image primitives such as edges, texture elements, or regions; intermediate level includes boundaries, surfaces and volumes; and high level includes objects, scenes, or events. There is no reason why there should be only three levels, and in fact, the task of transforming representations may be made easier by considering smaller jumps between representations. It should be clear in the descriptions that follow which level or levels are being addressed.

Representation of Prototypical Concepts. A prototype provides a generalized definition of the components, attributes, and relationships that must be confirmed of a particular concept under consideration in order to be able to make the deduction that the particular concept is an instance of the prototypical concept. A prototype would be a complex structure spanning many levels of description in order to adequately capture surfaces, volumes, and other events, to construct discrete objects into more complex ones, to define spatial, temporal, and functional relationships for each object, and to assert constraints that must be satisfied in order for a particular object in a scene to be identified.

Concept Organization. Three kinds of abstraction are commonly used, namely, feature aggregation, called "PART-OF", concept specialization, called "IS-A", and instantiation, called "INSTANCE-OF". The PART-OF hierarchy can be considered as an organization for the aggregation of concepts into more abstract ones or as an organization for the decomposition of concepts into more primitive ones, depending on which direction it is traversed. The leaves of the PART-OF hierarchy are discrete concepts and may represent image features. It should be pointed out that concept structure does not necessarily mean physical structure only, but similar mechanisms with different semantics may be used to also represent logical components of concepts. IS-A is a relationship between two concepts, one of which is a specialization of the other. An important property of the IS-A relationship is inheritance of properties from parent to child concept, thus eliminating the need for repetition of properties in each concept. Finally, the relationship between prototypical knowledge and observed tokens is the INSTANCE-OF relationship. These three relationships are typically used in conjunction with one another. Consideration of the semantics of these relationships is important, and such issues are discussed elsewhere (see INHERITANCE HIERARCHY).

Spatial Knowledge. This is perhaps the main type of knowledge that most vision systems employ. This includes spatial relationships (such as "above," "between," "left of"), form information (points, curves, regions, surfaces, and volumes), location in space, geometry, and continuity constraints (see REASONING, SPATIAL). Spatial constraints for grouping have appeared in the Gestalt literature in psychology and include the tendencies to group using smoothness of form, continuity of form, spatial proximity, and symmetry. The PART-OF relationship is used to represent aggregates of simple forms into more complex ones. Properties or attributes of spatial forms are also required, namely, size, orientation, contrast, reflectance, curvature, texture, and color. Maps are common forms of spatial knowledge representation, particularly for vision systems dealing with domains such as aerial photographs or navigation tasks.

Temporal Knowledge. Information about temporal constraints and time is not only necessary for the interpretation of spatio-temporal images but can also provide a context in which spatial information can be interpreted. Time can provide another source of constraints on image objects and events. Temporal constraints for motion groupings, in the Gestalt sense, include the tendencies to group using similarity of motion. The basic types of temporal information include time instants; durations and time intervals; rates, such as speed or acceleration; and temporal relations such as "before," "during," or "start." Each of these has meaning only if associated with some spatial event as well. PART-OF and IS-A relationships can be used for grouping and organizing spatio-temporal concepts in much the same fashion as for purely spatial concepts. A difficulty with the inclusion of temporal information into an IUS is that an implicit claim is made of existential dependency. That is, if a relationship such as "object A appears before object B" is included in a knowledge base, and object B is observed, then according to the knowledge base, it must be true that object A must have appeared previously (see REASONING, TEMPORAL).

The Scale Problem. It has been well understood since the early days of computer vision that spatial and spatio-temporal events in images exhibit a natural "scale." They are large or small in spatial extent and/or temporal duration with respect to scale, for example. This problem is different than the image resolution or coarseness problem, and there is no relationship between the two. It is important that an IUS deal with this as well. There are implications not only for the design of the image-specific operations that extract image events (a given operator cannot be optimal for all scales and thus is limited for a particular range of events that it detects well) but also for the choice of representational and control scheme. If spatio-temporal events require representation at multiple scales, the matching and reasoning processes must also be able to deal with the multiple scales. The unification of information from multiple scales into a single representation is important (see SCALE SPACE).

Description by Comparison and Differentiation. Similarity measures can be used to assist in the determination of other relevant hypotheses when matching of a hypothesis fails. This is useful in the control of growth of the hypothesis space as well as for displaying a more intelligent guidance scheme than random choice of alternates. The similarity relation usually relates mutually exclusive hypotheses. The relation involves the explicit representation of possible matching failures, the context within which the match failure occurred, binding information relevant to the alternative hypothesis, as well as the alternate hypothesis. Thus, the selection of alternatives is guided by the reasons for the failure.

Inference and Control Requirements

A brief note is in order before continuing this section on the difference between inference and control, particularly since in some works they are used as synonyms. Inference refers to the process of deriving new, not explicitly represented facts from currently known facts. There are many methods available for this task, and they are discussed in detail in other entries (see INDUCTVE INFERENCE; INFERENCE; REASONING). Control refers to the process that selects which of the many inference, search, and matching techniques should be applied at a particular stage of processing. The remainder of this section briefly discusses these issues and others in roughly the order that a designer of a typical image-understanding system would confront them.

Search and Hypothesis Activation. The basic interpretation paradigm used in IUSs, as is developed later in the Historical Perspective and Techniques section, is "hypothesize and test." There are several aspects to this, and these are described in turn beginning with search and hypothesis activation. A general vision system must contain a very large number of models that represent prototypical objects, events, and scenes. It is computationally prohibitive to match image features with all of them, and therefore, search schemes are employed to reduce the number of models that are considered. Only the salient models need be considered, and the determination of which are salient is termed the "indexing" problem. The catalog of search methods includes breadth-first, depth-first, hill climbing, best-first, dynamic programming, branch-and-bound, A*, beam search, information gathering or constraint satisfaction, relaxation labeling processes, and production systems. These are all described elsewhere (see A* ALGORITHM; SEARCH, BEAM; CONSTRAINT SATISFACTION; RULE-BASED SYSTEMS; SEARCH, BEST-FIRST; SEARCH, BRANCH-AND-BOUND; SEARCH, DEPTH-FIRST). A different categorization of search types, and one that is more frequently found in the IUS literature, is in terms of knowledge interactions. The following schemes are described below: model-directed search, goal-directed search, data-directed search, failure-directed search, temporally-directed search, hierarchical models, heterarchical models, blackboard models, and beam search. The choice of search method employed depends on a number of factors, including the form of the representation over which the search is to be performed, the potential complexity problems, and the goals of the search process.

Saliency of a model depends on the statement of goals for the search process. The search can be guided by a number of trigger features, for example, and any models that are encountered that embody those features are selected. The selection of a model for further consideration is termed "hypothesis activation." A search process that leads to a very large set of active hypotheses is not desired since the object of search is to reduce the space of models.

Matching and Hypothesis Testing. Once a set of active hypotheses has been determined, further consideration of each hypothesis takes place. The first task to be carried out is to match the active hypothesis to the data. It is important to note that data here do not necessarily only mean image-specific information. Matching is defined as the comparison of two representations in order to discover their similarities and differences. Usually, a matching process in vision compares representations at different levels of abstraction and thus is one of the mechanisms for transforming a given representation into a more abstract one. The result of a match is a representation of the similarities and differences between the given representations and may include an associated certainty or strength of belief in the degree of match.

The specific matching methods used depend largely on the representational formalisms that are used to code the data being compared. They can range from image–image matching, subgraph isomorphisms, or shape matching, to matching only selected features with a model, such as identifying structural components. Matching processes, particularly ones that involve matching images directly, are usually very sensitive to variations in illumination, shading, viewpoint, and 3-D orientation. It is preferred, therefore, to match abstract descriptions such as image features against models in order to overcome some of these problems. However, for 3-D models it is not always the case that image features can trigger proper models for consideration. Rather, the process must also involve the determination of the projection of the model that can be matched (see TEMPLATE MATCHING).

Generation and Use of Expectations. Expectations are beliefs as to what exists in the spatio-temporal context of the scene. The concept of expectation-directed vision is a common one that appears in most systems. Expectations must bridge representations in a downward direction, going from models to image appearance. *Projection* is a term commonly used to denote the connection between representations of the same concept but in differing domains. It is, for example, the relationship between a prototypical object and its actual appearance in an image. Thus, a mechanism is required that takes object position, lighting, observer motion, temporal continuity, and viewpoint into account to create an internal representation of an object's appearance in an image. Complete projections may not always be necessary, and in most cases it seems that expectations of important distinguishing features or structures are sufficient. The most common use of expectations is in directing image-specific processes in the extraction of image features not previously found (see also PARSING).

Change and Focus of Attention. Even the best of search and hypothesis activation schemes will often lead to very large hypothesis sets. Computing resources are always limited, and thus the allocation of resources must be made to those hypotheses that are most likely to lead to progress in the interpretation task. This can be done in a number of ways, including the use of standard operating system measures for resource allocation, as were used in an augmented fashion in HEARSAY (Erman and co-workers, 1980), ranking hypotheses by means of certainty or goodness-of-fit estimates, or by considering the potential of a hypothesis in conjunction with the expense that would be incurred in its evaluation. These best hypotheses, which are usually those that are confirmed or virtually confirmed, are also termed "islands of reliability."

Not only is it important to determine a focus of attention but it is also important to determine when to abandon a current focus as unproductive. The change of focus can be determined in one of two ways: the focus could be recomputed each time it was required or it could remain fixed and only change when circumstances necessitated the change. The latter is clearly more desirable; yet mechanisms for its implementation are few. It should be pointed out that a focus of attention does not necessarily refer only to a hypothesis set but may also refer to a region on an image or a subset of some representation.

One important type of perceptual attention is that referred to by the term "active vision" (see EARLY VISION; VISUAL RECOVERY). Active perceptual strategies, which provide for dynamic changes in the image acquisition process, are useful in at least the following ways: to see a portion of the visual field otherwise hidden; to compensate for spatial non-uniformity of a processing mechanism; to increase spatial resolution; to disambiguate aspects of the visual world (through induced motion, or lighting changes for example); to enhance the efficiency of processing by restricting the search space; and to provide for a better mathematical problem formulation. All of the above tacitly assume that some hypothesize-and-test mechanism is at work. Only if hypotheses are available, can a particular action due to an active perception mechanism actually yield benefits. Otherwise the search space is simply too large.

Certainty and Strength of Belief. The use of certainty measures in computer vision arose due to two main reasons: biological visual systems employ firing rate (which may be thought of as a strength of response), as the almost exclusive means of neural communication, and computational processes available currently are quite unreliable. This strength of response may be thought of as brightness for simplicity. Lateral inhibition (one of the processes of neural communication), whereby neurons can inhibit the response of neighboring ones based mainly on magnitude of the firing rate, is a common process, if not ubiquitous. It motivated the use of relaxation labeling processes in vision. In relaxation, the strength of response is termed "certainty," and is often used as a measure of reliability of a corresponding decision process, for example, the goodness of fit of a line to the data. Since visual data are inherently noisy due to their signal nature, measures of reliability are important in the subsequent use of information derived using unreliable processes.

Yet another use of certainty is in hypothesis ranking. The ranking of hypotheses is useful not only for the determination of a focus of attention but also for determining the best interpretation. Most schemes introduce some amount of domain dependence into the control structure, and this seems to lead to problems with respect to generality. An important problem is the combination of certainties or evidence from many sources.

Inference and Goal Satisfaction. Inference is the process by which a set of facts or models is used in conjunction with a set of data items to derive new facts that are not explicitly present in either. It is also called reasoning. The many forms of reasoning include logical deduction, inheritance, default reasoning, and instantiation. (See INHERITANCE HIERARCHY; REASONING, DEFAULT.)

However, it should be pointed out that the vision problem adds a few different wrinkles to this task that may not appear in many other reasoning processes. It is not true in general that the data set is complete or correct, and processses that can reliably draw inferences from incomplete data are required. Second, since vision is inherently noisy and as described above requires reliability measures, inference schemes should also permit reliability measures to be attached to derived conclusions. Finally, since the process of vision involves a transformation from images to a final description through many intermediate representations, a reasoning scheme must be able to cross between several representations.

Most IUSs are not explicitly driven by a goal when interpreting images. They typically have implicit goals, such as to describe the scene in terms of volumetric primitives, to describe everything in as much detail as possible, or to describe the scene in the most specific terms possible. Human vision usually does involve a goal of some kind, and the area of AI that is concerned with how to achieve goals given a problem is called "planning." Systems that can plan an attack on a problem must contain metaknowledge, that is, knowledge about the knowledge that the system has about the problem domain (see META-KNOWLEDGE, META-RULES, AND META-REASONING). The metaknowledge allows the system to reason about its capabilities and limitations explicitly. Such systems have a set of operations that they can perform, and they know under which circumstances the operations can be applied as well as what the effects may be. In order to satisfy a goal, a sequence of operations must be determined that, in a stepwise fashion, will eventually lead to the goal. Attempts to find optimal plans usually are included in terms of minimization of cost estimates or maximization of potential for success. In vision the sequence of operators may involve image feature extraction, model matching, and so on (see PLANNING).

HISTORICAL PERSPECTIVE AND TECHNIQUES

The historical development of the techniques of image understanding provides an interesting reflection of the major influences in the entire field of AI. The emphasis in the IU community has been primarily in the control structure, and this discussion begins with the sequence of contributions that led to the current types of control mechanisms. Rather, little emphasis has been placed on integrating the best of the early vision schemes into IUSs, and one notices the range of weak solutions to the extraction of features. Little discussion is thus provided; however, in the description of control structures for specific systems, appropriate notes are made.

Control Structures

The heart of virtually all IU systems is the control structure (qv). Features universal to all working IUSs are cyclic control involving feedback (see CYBERNETICS) and the requirement of specific solutions to the problem of uncertainty. This survey of the development of control structure highlights only those systems that require and use explicit models of objects or events of the domain. Other important contributions that impact IUSs are allocated their appropriate historical due but are not considered part of the direct line of development. Finally, with two exceptions, the hypothesis of Marr and Nishihara (1978) and the intrinsic image concept of Tenenbaum and Barrow (1977), only implemented and tested systems are described in this section.

Developing the Cycle of Perception. Roberts was the first (1965) to lay out a control scheme for organizing the various components of a vision system. They are shown pictorially in Figure 1. He defined several of the major processing steps now found in all vision systems: extract features from the image, in his case, lines; activate the relevant models using those features; project the model's expectations into image space; and finally, choose the best model depending on its match with the data. This is not a true cycle, and because of the lack of feedback, it was very sensitive to noisy input. Falk (1972) realized that Roberts' work involved an assumption that would rarely be satisfied in real application domains, namely, that of noise-free data. If noisy data were to be correctly handled, enhancements to Roberts' processing sequence were required (1972). In Figure 2 Falk adds a new component, the fill in incompleteness step, and closed the loop, allowing partly interpreted data to assist in the further interpretation of the scene. His program was called INTERPRET.

Shirai (1973) defined a system for finding lines in blocks world scenes and interpreting the lines using models of line junctions and vertices for polyhedral objects. Thus, he was able to use interpreted lines as guidance in subsequent line finding. He first extracted features from a reduced image, thus smoothing out some of the noise and smaller detail features, and then used these gross features in subsequent guidance. Shirai's cycle is shown in Figure 3. Shirai, however, was not the first to employ reduced images in a preprocessing stage. Kelly

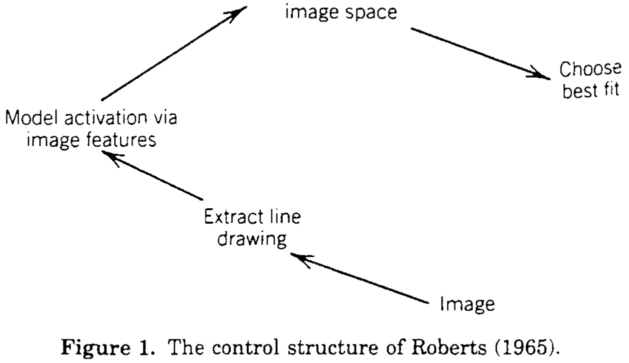

Project models into image space
Choose best fit
Model activation via image features
Extract line drawing
Image

Figure 1. The control structure of Roberts (1965).

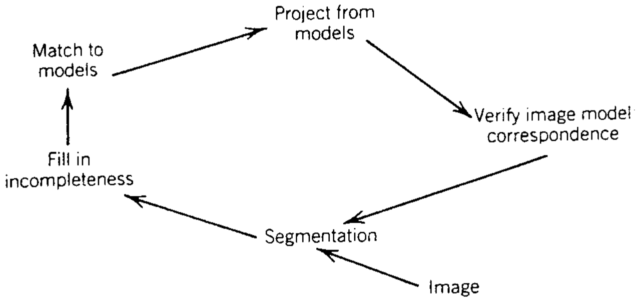


Figure 2. The control structure of Falk (1972).

(1971) had the intuition that if an image that was reduced in size was processed initially, instead of the full-size image, much of the noise could be reduced, and the resulting edges of lines could be used as a plan for where to find edges and lines in the full image. This was applied to the domain of face recognition. Kelly reduced an image to 64 × 64 pixel size, thus minimizing noise effects, and then located the outlines of the faces. Those outlines then formed a plan for the full-size image, limiting the search space for the detailed facial outlines. However, Kelly's system contained no models and was a sequential two-step process.

Several incarnations of the cycle appeared subsequently, and one example of note is presented here, namely, the work of Tenenbaum and Barrow (1977) in their interpretation-guided segmentation (IGS) program. Their version of the cycle is shown in Figure 4. IGS experimented with several types of knowledge sources for guidance of the segmentation process: unguided knowledge, interactive knowledge, both user driven and system driven; models; and relational constraints. They concluded that segmentation is improved with the application of knowledge when compared to the unguided case, and with little computational overhead—the more knowledge, the faster the filtering process. Perhaps the most elegant portrayal of the cycle of perception, and also the coining of the term itself, is due to a contribution by Mackworth (1978) and is shown in Figure 5. This basic cycle appears, in a variety of forms, in virtually all IUSs that have appeared since. Kanade's modification of the cycle (1980) explicitly included the separation of scene domain and image domain considerations, a requirement that was first pointed out by Huffman (1971) and also independently by Clowes (1971). This refers to the difference between an object's 2-D appearance in an image versus an object's 3-D representation in the world. Figure 6 portrays Kanade's cycle.

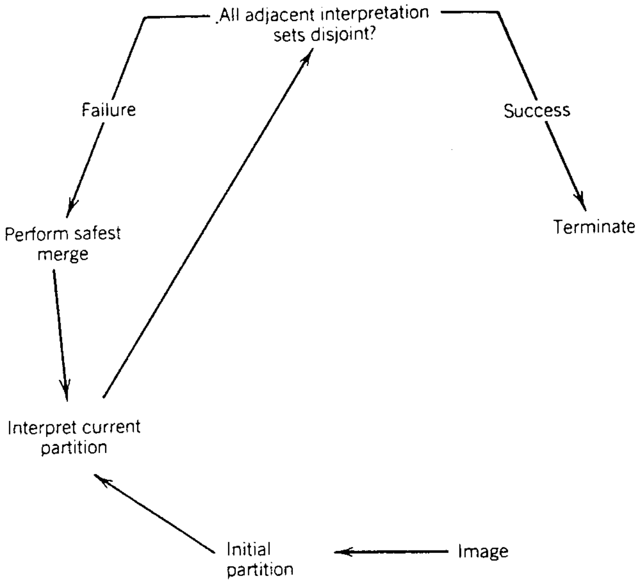

All adjacent interpretation sets disjoint?
Failure
Success
Perform safest merge
Terminate
Interpret current partition
Initial partition
Image

Figure 4. The interpretation-guided segmentation control structure of Tenenbaum and Barrow (1977).

Tsotsos and co-workers (1985) further elaborated the model for the ALVEN system by specifying exactly at which points of the cycle the different hypothesis activation (or indexing) methods are applied. In addition, since his task was to understand visual motion, the element of time was also added. To this point in the development of the cycle of perception, although use had been made of different representational tools for organizing models, no explicit consideration had been given to how to best take advantage of the organization. Tsotsos used the common organizational tools of specialization (IS-A), decomposition (PART-OF), and SIMILARITY (mutual exclusion of models, or winner take all) and add temporal precedence in order to organize a large set of models. His cycle is

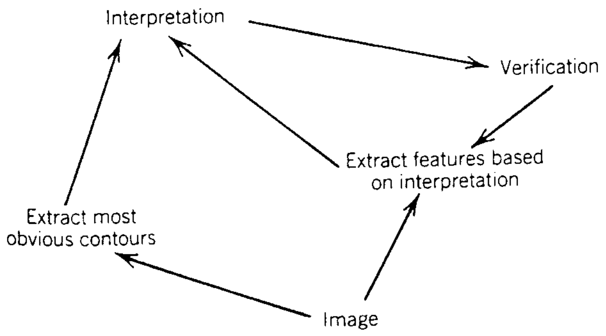

Interpretation
Verification
Extract features based on interpretation
Extract most obvious contours
Image

Figure 3. The control structure of Shirai (1973).

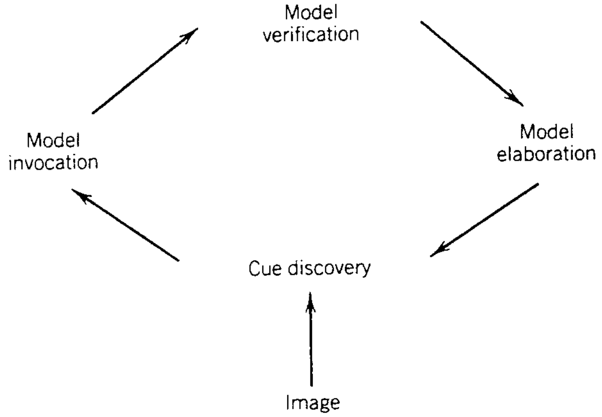


Figure 5. The cycle of perception of Mackworth (1978).

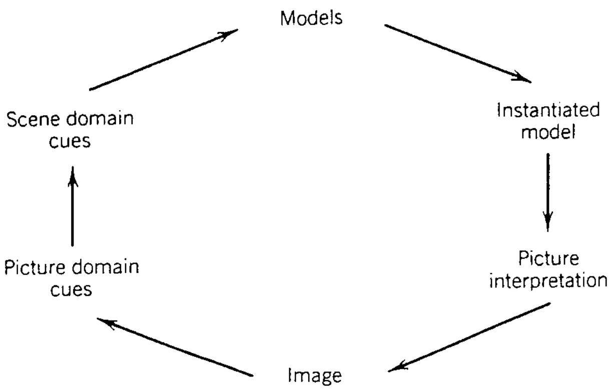


Figure 6. The control structure for Kanade (1980).

shown in Figure 7. Definitions of the different hypothesis activation methods driven by knowledge organization relationships were also provided by Tsotsos. The methods are briefly summarized below.

Goal-Directed Activation. The goal of the vision system is to find the most specific, or specialized, description in the system's repertoire for the image contents. The specialization of hypotheses involves top-down traversal, from general to specific, on an IS-A hierarchy, moving downward when concepts are verified. Verification of an IS-A parent concept implies that perhaps one of its IS-A children applies, although the confirmation of a concept implies that its IS-A parents must also be true. Multiple IS-A children can be activated, but a more efficient scheme would be to activate one of the children if all children form a mutually exclusive set, or one from several such sets, and then allow failure-directed search to take over.

Model-Directed Activation. The elaboration of models involves top-down traversal of the PART-OF hierarchy. This too implies a constrained form of hypothesize and test for components of classes that reflect greater resolution of detail. Movement down the PART-OF hierarchy forces activation of hypotheses corresponding to each of the components of the PART-OF parent hypothesis.

Data-Directed Activation. The PART-OF hierarchy can also be traversed bottom up in aggregation mode. Bottom-up traversal implies a form of hypothesize and test, where hypotheses activate other hypotheses that may have them as components.

Failure-Directed Activation. Failure-directed search is along the SIMILARITY dimension. Typically, several SIMILARITY links will be activated for a given hypothesis, and the resultant set of hypotheses is considered as a discriminatory set, that is, at most, one of them may be the correct one. SIMILARITY interacts with the PART-OF relationship in that exceptions raised that specify missing components are handled by the hypothesis' PART-OF parent, the hypothesis that contains the context within which the exception occurred.

Temporally Directed Activation. Temporal search is a special case of model-directed search along the PART-OF dimension. Concepts may represent compound temporal events, such as sequences, simultaneous events, or overlapping events. In a sequence each element of the sequence has a PART-OF relationship with the event. Thus, on activation of the class, it is meaningless to activate all parts, as stated above, at the same time. Activation of the parts only occurs when their particular temporal specifications are satisfied. Temporally-activated hypothesis invocation and prediction is an instance of an active vision

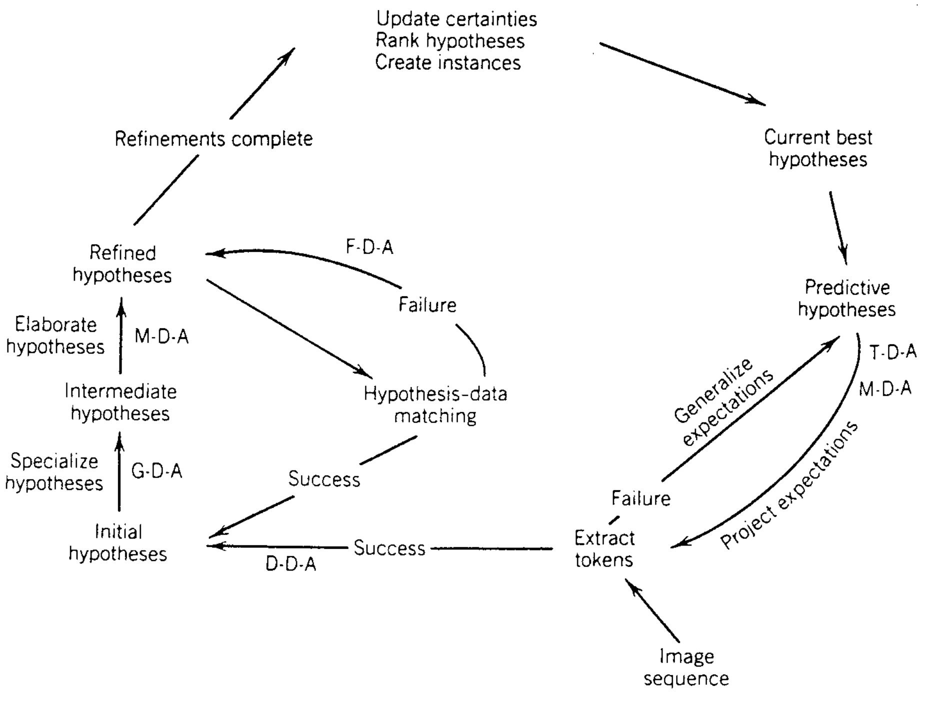


Figure 7. The control cycle for motion understanding of Tsotsos (1985). Hypothesis activation types: DDA, data-directed activation; GDA, goal-directed activation; FDA, failure-directed activation; TDA, temporally directed activation; MDA, model-directed activation.

strategy. This necessarily requires time and image samples acquired over time.

Marr, usually credited with contributions only in early vision, also had specific processes in mind for the high levels of vision presented in his book *Vision* (1982). It would indeed have been interesting to have seen an attempt at implementation and testing of his ideas. Marr, in his own words, viewed "recognition as a gradual process that proceeds from the general to the specific and that overlaps with, guides, and constrains the derivation of a description from the image." He proposed that a catalog of models be constructed using volumetric primitives and organized using a specialization hierarchy (IS-A) as well as a decomposition hierarchy (PART-OF). Models were selected based on the distribution of components along principal axes of the derived volumetric primitives represented in the 3-D sketch. He proposed three indexing schemes: The primary one was the "specificity index," traversal from general to specific models (goal-directed); the secondary ones, used in support of the first, were the "adjunct index," traversal from models to model components (model-directed), and "parent index," traversal from model components to parent models (data-directed). The model provided relative orientation constraints used to determine absolute orientation. An image space processor then related image-centered and object-centered descriptions and computed relative lengths of component axes. This new information can be used to disambiguate shapes at the next level of specificity. It is interesting to note that Marr did not propose a cycle of processing and that the 3-D sketch represented all possible information derivable directly from the image. In general, this is not realizable, and a scheme without feedback is insufficient.

In several models the issue of feedback and the relationship between explicit models and their appearance in an image was mentioned. The projection of hypotheses into image space is a difficult problem for which few solutions exist. As pointed out previously, expectations have been used in most IUSs since Kelly's and Shirai's work. Expectations were used in the SEER system of Freuder (1977) to guide region growing and identification of specific portions of a hammer. A thorough understanding of human body motions and a model of the allowed joint configuration enabled the design of a constraint propagation network that integrated current motions and known body positions with hypothesized ones, producing expected locations in 3-D for given body joints (O'Rourke and Badler, 1980). An interesting conclusion from the ALVEN system's use of expectations is that the information contained in an IS-A hierarchy of concepts can be exploited for the generation, verification, and modification of expectations of actual object appearance in a sequence of images. If expectations fail, movement up the hierarchy to a more general concept provides the next best alternative consistent with the semantics of the interpretation. However, the key problem of relating 3-D object viewpoint-independent models to image-specific ones is still an outstanding one. A good example of work on this topic is the ACRONYM system (Brooks, 1981). Given a geometric object model and viewpoint and illumination, ACRONYM predicts partial object appearance in the image. That is, only the important features required for identification are predicted since the whole problem is so computationally expensive. Recently, an additional dimension to this problem has been added through active strategies. Califano, Kjeldsen, and Bolle (1990) provide an interesting approach which uses models combined with foveation strategies in order to deal with feedback, tractability, and integration of successive sensor fixations.

Heterarchical Models. A heterarchical model of vision is one made up of a collection of separate modules, each module performing some specialized task and each communicating with all others as appropriate. Freuder was perhaps the first within the vision community to apply such an idea in his system for recognizing tools called SEER (1977). "Active knowledge" was his term for the use of procedural knowledge in directing the control; Freuder's work is thus an early precursor to the recent wave of interest in active perception strategies. Knowledge was represented as semantic networks (qv). Nodes represented objects and links represented how objects help establish one another. Each object encoded procedural knowledge, and together the objects formed the set of modules, each communicating with other relevant modules.

Another form of heterarchy is the "demon" scheme, where each knowledge source continuously monitors a database of assertions about the images and of models to see if its prerequisites are present. If found, the demon then carries out some actions that may involve changes to the database. Badler (1975) used a demon model for event analysis, and each demon represented the knowledge required to recognize a particular event type. Two other specific versions of heterarchy are presented by Nevatia (1978) and Levine (1978). They provide two other views for the composition of the collection of modules. They are presented in Figures 8 and 9. Perhaps the main conclusion that can be drawn from the heterarchical models is that as the number of interacting modules grows, the communication and organization problems increase dramatically.

Hierarchical Models. Hierarchical models are comprised of a specialized collection of modules, but the communication pathways are restricted, reflecting an ordering of both processing steps and levels of abstraction in

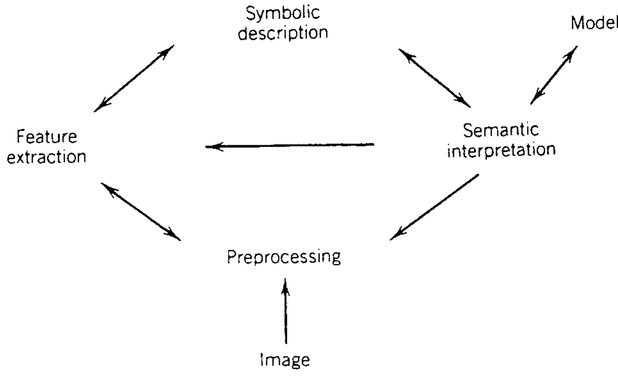


Figure 8. The control structure of Nevatia (1978).

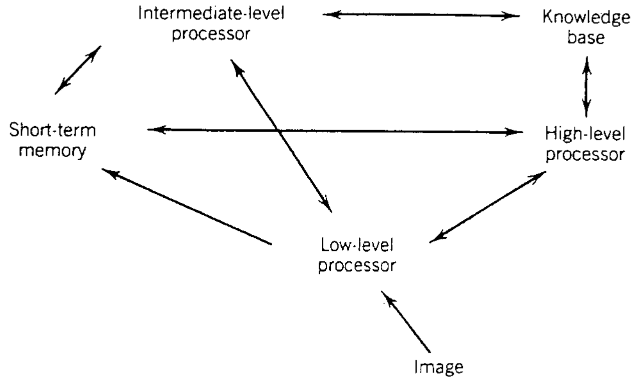


Figure 9. The control structure of Levine (1978).

the computation. One of the best known is due to Barrow and Tenenbaum (1978) and is diagrammed in Figure 10. This model reflects a major contribution in representation, namely the idea of "intrinsic images." This is described in Spatial Relationships, below.

Another important hierarchical model that elaborates on Barrow and Tenenbaum's model is that of the VISIONS system (Hanson and Riseman, 1978). This fills in several details regarding communication and control across the multiple levels of representation that are present in all image-understanding systems. Yet another specific type of hierarchical model emerged, conforming with the basic definition and philosophy but also attempting to provide a solution to the spatial scale problem. Uhr called these models "recognition cones" in his contribution (1972), and they have also been termed "pyramid models" (Tanimoto and Klinger, 1980). The major distinctions come from the facts that each layer of the cone computes image properties at successively coarser resolutions and each computation communicates only with computations occurring in layers immediately above or below or with computations within the layer. An unfortunate result of this idea is the linking of spatial scale with resolution; as noted earlier, the optimal scale for the detection of specific spatial forms has little relationship to image resolution.

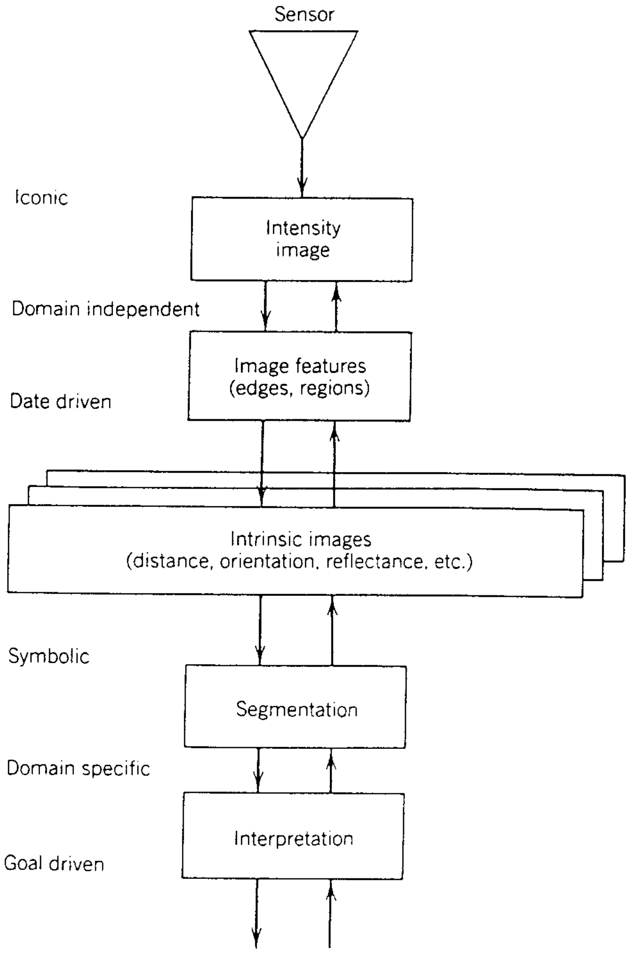


Figure 10. The control structure of Barrow and Tenenbaum (1978).

Blackboard Models. Blackboard models (see BLACKBOARD SYSTEMS) were borrowed for use in vision from the HEARSAY work in speech understanding. In fact, they are a specific form of heterarchy in that each knowledge source (module) can communicate with any other. Knowledge sources are organized hierarchically. The major difference and improvement over the versions of heterarchy that were presented earlier is that the communication occurred through a global data structure called a blackboard rather than the communication pathways being fixed. The VISIONS system (Hanson and Riseman, 1978) incorporates this idea as well as pyramid processes. The knowledge sources defined are inference net, 2-D curve fitting; 2-D shape; occlusion; special attribute matcher; 3-D shape; perspective; horizon; and object size. The VISIONS structure is shown in Figure 11. The advantages of blackboard models include their modularity; however, their utility in speech has not been repeated in vision, primarily because of the important differences between speech and vision.

Beam Models. Once again, speech understanding influenced the design of a vision system. In this case the HARPY system (Lowerre and Reddy, 1980) influenced the 1980 design of the ARGOS system of Rubin (1980). Rubin's work is interesting because it was the only attempt to use beam search (qv) (also called locus search) in vision. Beam search produces a "beam," a pruned search tree that contains a list of near-miss alternatives around the best path. Both signal and model characteristics are included in this consideration. The scheme as realized in ARGOS is not one that has promise for general-purpose vision systems. ARGOS looked at images of downtown Pittsburgh, attempting to classify regions as sky, buildings, or mountains, for example. The network over which the beam search was performed was a large one whose nodes were pixels or image regions and whose arcs were spatial relations.

Rule-Based Approaches. Rules (of the if <premise> then <action> form) were introduced into vision at about the same time that they appeared in production systems. The introduction is due to Baird and Kelly (1974), who

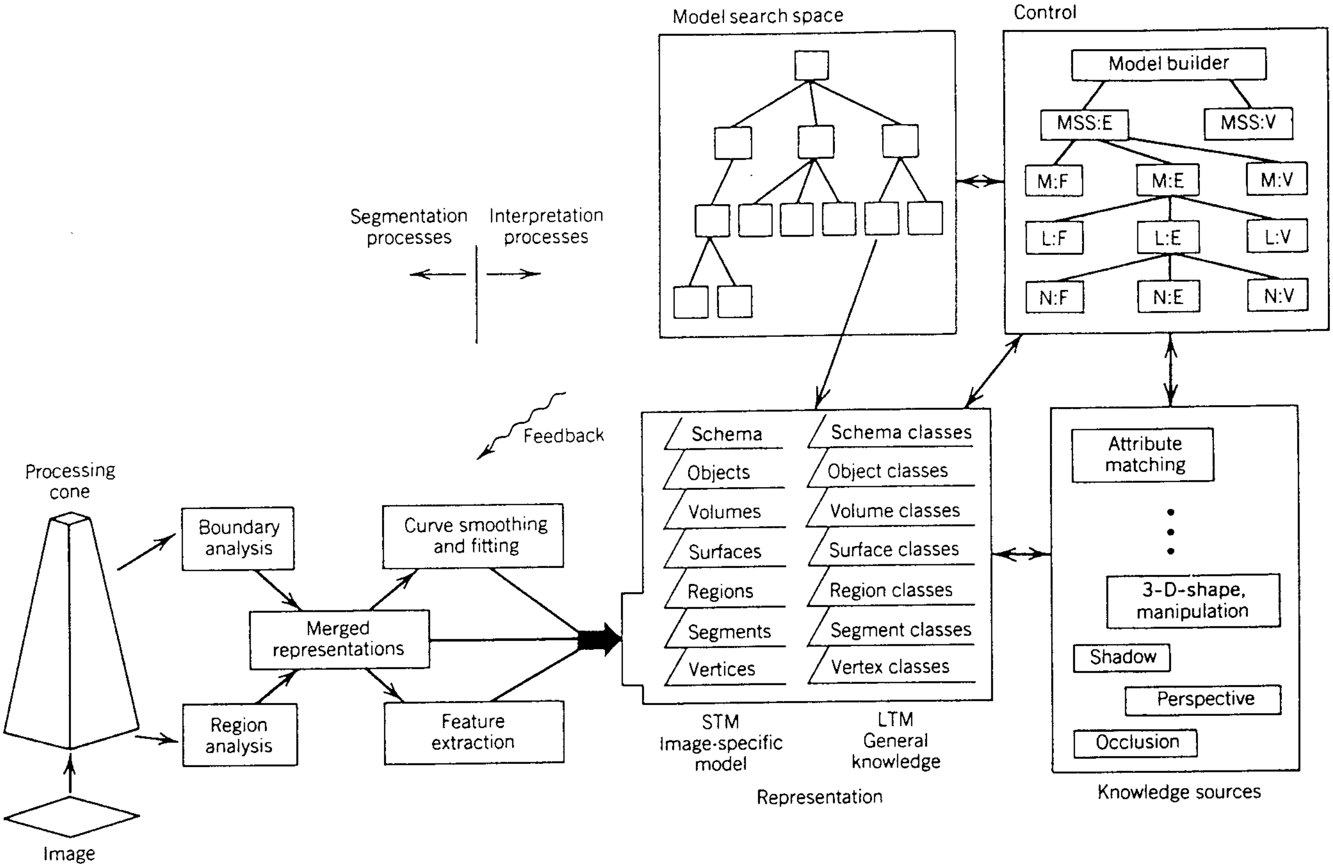


Figure 11. The Blackboard structure of VISIONS (Hanson and Riseman, 1978).

claimed that context is a necessary consideration in the development of their paradigm for semantic picture recognition. They used inference rules to incorporate contextual considerations, and premises were features extracted from images. More recently, perhaps due to the success of the expert systems approach, several other vision systems have appeared that utilize rule-based knowledge and reasoning (see RULE-BASED SYSTEMS).

Typically, pure data-directed reasoning is insufficient as described above, and rules are fired in both goal-directed (backward-chaining) and data-directed (forward-chaining) modes (see PROCESSING, BOTTOM-UP AND TOP-DOWN). Rules are used to represent various facts about images. For example, in SPAM, the system of McKeown and coworkers (1984), rules are used to encode spatial relationships among entities in the scene as well as to encode constraints on sizes and shapes of visual entities. Rule-based reasoning is used to provide the system with the best next task based on the strength of expectations as well as for the generation of expectations. Other IUSs that employ rule-based reasoning are the systems of Nagao and Matsuyama (1980); Ohta (1980); Ferrie, Levine, and Zucker (1982); and Riseman and Hanson (1984).

Representation Formalisms

The development of representational tools used in the IUS community mirrors quite closely developments in other subdisciplines of AI. The use of heuristics (qv) reflects the power-based era of AI. The appearance of semantic networks (qv) in the memory-modeling community and their use by the knowledge representation (qv) and language-understanding communities (see NATURAL LANGUAGE UNDERSTANDING) influenced their use in IUSs. Blackboards and beam searches were developed for the major speech-understanding systems (HEARSAY and HARPY, respectively) and subsequently appeared in vision systems. Minsky's (1975) frame theory (see FRAMES), developed with a specific eye toward vision, was used in several vision systems. The success of expert systems prompted the use of rule-based approaches in IUSs as well.

Spatial Representations. Vision systems require the explicit representation of points, curves, surfaces, and volumes. There are a number of schemes that are employed, namely, points, line segments, splines, fractals, and generalized cylinders, among others. As an example, the VISIONS system employs a representation of 3-D complex surfaces and 2-D curves based on B-splines and surface patches and also makes use of the PART-OF and INSTANCE-OF relationships in building complex structures. ACRONYM uses a generalized cylinder representation in conjunction with PART-OF and IS-A organizations. There is no real consensus yet on what constitutes an adequate set of primitives for spatial representations. Discussions and examples can be found for several representational points of view: Marr and Nishihara (1978) for generalized cylinders; Pentland (1985) for the super-quadric approach; Kass, Witkin and Terzopoulos for snakes (1988); Terzopoulos, Kass and Witkin for deformable, symmetry-seeking 3-D models (1988); and Biederman (1988) for geons, etc (see SHAPE).

Much work in representation and reasoning about space has appeared outside the vision community. Comparison of object location and the representation of the corresponding relations is considered in Freeman (1975). Kuipers (1978) describes his TOUR model for route-solving problems and discusses the spatial knowledge relevant to that task. McDermott and Davis (1984) also include a representation for spatial knowledge and a scheme for reasoning about it. However, both Kuipers and McDermott and Davis were concerned with spatial route-finding tasks, and this is not directly comparable to the reasoning required for vision systems. Spatial representations are covered in other entries (see REASONING, SPATIAL). The representation of maps is quite straightforward and does not require further elaboration. The interested reader should consult articles on the MAPSEE (Mackworth and Havens, 1983) or HAWKEYE (Barrow and co-workers, 1977) systems.

Two specific representations can be considered major contributions, namely, the schemes of Marr (1982) and Barrow and Tenenbaum (1978). Marr proposed a progression of representations that he termed the "primal sketch," the $2\frac{1}{2}$-D sketch, and the 3-D sketch. The primal sketch represented information about the 2-D image, primarily intensity changes and their organization. The $2\frac{1}{2}$-D sketch represented the orientation and depth of surfaces and discontinuity contours. Finally, the 3-D sketch represented shapes and their spatial organization in an object-centered manner. In contrast, Barrow and Tenenbaum claimed that the appropriate intermediate-level representation consisted of a number of separate feature maps, all image centered, that perhaps interact in order to be computed unambiguously. These features include surface discontinuities, range, surface orientation, velocity, and color.

Heuristics. The use of heuristics (qv) appears in most vision systems in one form or another. Systems that used only heuristics, however, appeared only during the power-based era of AI and do not really qualify as IUSs using the definition requiring explicit object or event models. Those systems typically deal with blocks-world scenes.

Semantic Networks. Semantic networks (qv), that is, graph structures whose nodes represent objects or events and whose arcs represent relationships between the objects and events, have made an important impact on IUSs. Two examples are the work of Levine (1978) and that of Badler (1975). Levine's system deals with the interpretation of natural scenes, and he constructs a knowledge base with nodes representing entities such as sky, road, and house. Arcs represent spatial relations, such as left of, above, or behind. Badler used the same idea but represents events as well as objects with nodes, whereas arcs represent spatial as well as temporal relations.

Frames. Minsky's frame theory (1975) was one of the most influential works within the representation community, and since it was designed as a representation for vision, it left a mark on the IUS community as well. Frames are data structures representing a prototypical object or event. The components of the structure are slots that are filled with specific instances of visual entities. Slots may specify a type of instance, may specify a default value that can be used if the instance is not found, and may have associated constraints that relate one slot to others. Frames, sometimes also called "schemata," are used in the SIGMA (Matsuyama and Hwang, 1985), ALVEN (Tsotsos, 1985), ACRONYM (Brooks, 1981), MAPSEE (Mackworth and Havens, 1983), and VISIONS (Hanson and Riseman, 1978) systems among others.

The concept of a "representation space" was described by Bobick and Bolles (1989) in an attempt to deal with the integration of visual information over time as it was acquired. This space is a lattice of evolving representations as the certainty in an object's description increases as more data is acquired. This represents early work on an important issue.

A large collection of frames poses a serious indexing problem, and one solution for this is to organize the frames into a semantic network. In such a representation nodes are frames that represent objects or events, and arcs are network organizational primitives, such as generalization, specialization and aggregation, decomposition. The similarity relationship, motivated by Minsky, is added to the ALVEN scheme, as well as a temporal precedence dimension, as further organizational relations among frames.

Rules. Rules may be used to encode object characteristics, spatial relationships among objects, constraints on shape and sizes, and so on, for use in an IUS. The use of rules in the SPAM system (McKeown and co-workers, 1984) has already been mentioned. In the VISIONS system (Parma and co-workers, 1980) rules are applied to the attributes of the lines and regions in an intermediate representation. Simple rules define ranges over a feature value and, if fired, are considered as a vote for an object label. Here image features include color, texture, shape, size, and location, and feature values include length, location, orientation, contrast, and width. They allow complex combinations of simple rules. For example, they have a rule that measures excess green present in grass by computing the appropriate mean of color values in the R-G-B ranges for pixels in the region in question. This approach can also be found in the work of Nagao and Matsuyama (1980), and Ohta (1980).

Reasoning and Uncertainty

Relaxation-Labelling Processes. Relaxation-labeling processes appeared first as discrete constraint propagation schemes and then as probabilistic ones (see REASONING, PLAUSIBLE). The primary difference between the discrete and continuous schemes is that decisions in the discrete case are binary—a label is either true or it is removed from consideration—and in the continuous case, labels have an associated strength that is increased or decreased depending on the constraints imposed on it by its neighboring context. One may think of strength in this context as a measure of goodness of fit—it is not a probability in the formal sense (see Hummel and Zucker, 1980, on continuous relaxation). Relaxation labeling is commonly used in recognition cone approaches, within layers of the cone, and hierarchically between layers. Also, the excursion into a time-varying continuous relaxation scheme called "temporal cooperative computation" is presented by Tsotsos (1987).

Evidential Reasoning. One method for making decisions based on uncertain information is the use of Bayesian probabilities. This method is described elsewhere (see BAYESIAN INFERENCE METHODS). Another method for combining evidence in order to draw conclusions that has been applied in vision is the Dempster-Shafer Theory (qv). The major difference between Dempster-Shafer and Bayesian probabilities is that an explicit representation of partial ignorance is provided. Belief is represented in the range [0,1], and lower bounds within this interval are moved higher and upper bounds are moved lower, reflecting the addition of supporting or conflicting evidence, respectively. The width of the remaining interval is regarded as ignorance. This scheme is being applied in the VISIONS work (Parma and co-workers, 1980).

Jepson and Richards (1990) have considered perceptual reasoning given beliefs of the world. They have proposed a "lattice theory" where a lattice of belief states reflects the possible interpretations of a scene. A local maximum in the lattice partial ordering is the requirement for an acceptable percept.

Spatiotemporal Reasoning. Reasoning systems that deal primarily with axioms whose propositions are spatial relations or facts can be termed "spatial reasoners." Similarly, those dealing with temporal relations are "temporal reasoners" (see REASONING, TEMPORAL), and those that deal with geometric information are "geometric reasoners." Grouping processes, such as those reflected by inferences along the PART-OF representation dimension, are also included here. It is clear that the inclusion of such reasoning processes is important in IUSs. ACRONYM (Brooks, 1981) uses 3-D object models and can reason about complex coordinate transforms of them. It also includes an algebraic reasoner that reasons about sets of non-linear algebraic symbolic inequalities and bounds and determines satisfying sets for those inequalities. Other systems that explicitly address the problem of spatial reasoning are SIGMA (Matsuyama and Hwang, 1985) and SPAM (McKeown and co-workers, 1984). In both cases, the reasoning is 2-D and is based on image-centered representations. A specific type of spatial reasoning uses maps. The premise behind the use of maps is that explicit map-to-image correspondence can be derived using models of the imaging process and models of the terrain in the maps. The correspondence can be used to guide the interpretation of detailed features of the image (See REASONING, SPATIAL.)

An example of an IUS that deals with temporal reasoning is the ALVEN system (Tsotsos, 1985). The form of reasoning is very different than the temporal calculus of Allen (1984), which is an example of the pure temporal reasoning methods. Allen's scheme was not intended for vision, and it therefore displays several deficiencies that are important for vision: it does not allow for strength of belief in a temporal relation; it does not provide a recognition structure for detecting and labeling temporal relations; and it does not account for the fact that in a real-time recognition situation, all data in time are not available to the system. The ALVEN framework incorporates all of these points, in addition to the fact that all temporal relations in ALVEN are really spatiotemporal.

Planning. As mentioned previously, planning (qv) has played a role in vision since Kelly first used plans in his program for face recognition (1971). Kelly applied edge operators to a reduced image in order to extract the face outline and then expanded the outline to the original image size and searched for details only within this prediction window. This type of planning, using explicit prediction windows, has been used in many systems.

An example of an IUS that uses planning with goal satisfaction is Garvey's system (1976). In the domain of indoor office scenes, Garvey defined operators such as "find seat" (of a chair), "validate seat," "grow seat," and similarly for all objects that were known. Sequences of operators were planned and represented in an AND/OR tree (see AND/OR GRAPHS). Plans were scored depending on cost and confidence. On execution, the outcome of particular steps can be used to modify other parts of the plan. The system of Ballard, Brown, and Feldman (1978) also has a limited planning capability. It is limited in that only a very small number of operators are available, and no plan hierarchy is constructed. In the domain of locating ribs in chest radiographs, for example, Ballard and co-workers included three independent rib-finding procedures that were managed by an executive procedure.

EXAMPLE SYSTEMS FOR SPECIFIC PROBLEM DOMAINS

The description of systems provided in Table 1 is necessarily abbreviated and incomplete. It does not include all systems, nor all details for each system included. The table presents information for each system, along with relevant pointers to the literature. All systems employ the basic cycle of perception in some form, perhaps with important enhancements that have been described previously, unless otherwise noted. Thus, they all involve the interaction of both top-down and bottom-up methods. All systems make the assumption that knowledge can compensate for poor quality input and weak image-specific segmentation processes. All systems have demonstrated some reasonable level of performance, usually on a small set of carefully chosen example images. The systems are grouped according to application domain and are listed alphabetically by the system name or the principal author's name. Within each category, at least one example of

Table 1. Example Systems for Specific Problem Domains

Name	Authors	Institution	References	Domain	Representation	Control
Aerial Photographs						
ACRONYM (see Figure 12)	T. Binford R. Brooks R. Greiner	Stanford University	Brooks, 1981; 1983; Brooks and co-workers, 1979	Airport scenes	3-D geometric models; generalized cones; ellipses, ribbons; frames; PART-OF; IS-A; object graphs for geometric constraints; restriction graphs for algebraic constraints; context graph; coarse-to-fine detail; models independent of viewpoint; user interface for model definition using volumetric primitives	Line finding; rule-based problem solving; graph matching between prediction, graph, and picture; graph of image features; prediction of object appearance based on viewpoint and illumination, but only of important features; geometric reasoning; algebraic reasoning
Not available	D. Ballard C. Brown J. Feldman	University of Rochester	Ballard and co-workers, 1978	Ship dock scenes in satellite photos	2-D spatial knowledge; semantic network; meta-knowledge for planning; sketch map as intermediate image representation; procedural knowledge	Distributed control; model-image mapping via procedural knowledge of objects; executive chooses most likely mapping procedure
HAWKEYE	H. Barrow R. Bolles T. Garvey T. Kremers J. Tenenbaum H. Wolf	SRI International	Barrow and co-workers, 1977	Aerial photographs	2-D topographic maps as symbolic scene model; geometric camera model	Parametric correspondence for map matching camera model calibrated on landmarks, then used to predict precise locations of other features
Not available	W. Cole A. Huertas R. Nevatia	University of Southern California	Huertas and co-workers, 1989	Airport scenes	2-D spatial knowledge; generic knowledge about airport and associated structures; structures represented by their parts and boundaries	Hierarchical hypothesize-and-test; perceptual grouping for initial hypotheses; objects decomposed into parts for image search directed by hypotheses
MAPSEE-1, MAPSEE-2, MAPSEE-3	W. Havens A. Mackworth J. Mulder	University of British Columbia	Mackworth, 1977; Mackworth and Havens, 1983; Mulder and co-workers, 1988	Freehand drawings of maps on satellite images	2-D spatial knowledge; cartographic elements; schemata; IS-A; PART-OF; Waltz-like primary cues in drawings such as TEE, OBTUSE L, MULTI; composition and specialization hierarchies; discrimination graphs	Extended Waltz filtering to n-ary relations and hierarchies (hierarchical arc consistency); region growing
Not available	T. Matsuyama M. Nagao	Kyoto University	Nagao and Matsuyama, 1980	Aerial photographs of roads, houses, forests, fields, and rivers	2-D spatial knowledge; regions with attributes including spectral information; objects defined using 2-D heuristics	Blackboard-style specialized subsystems for specialized features; interpretation is image centered
SIGMA	V. Hwang T. Matsuyama	Kyoto University, University of Maryland	Matsuyama and Hwang, 1985	Aerial photographs of roads, houses, forests, fields, and rivers	Frames; PART-OF; IS-A; rules attached to slots for constraint and instantiation information; 2-D spatial knowledge; spectral knowledge	Three communicating experts, geometric reasoner, model selector, and low level; intersection of prediction areas in image-centered representation; evidence accumulation in image-centered representation
SPAM (see Figure 13)	W. Harvey J. McDermott D. McKeown	Carnegie Mellon University	McKeown, Harvey, and McDermott, 1984	Airport scenes	2-D spatial knowledge; pyramids for image features; viewpoint dependent; short- and long-term memory; relational database	Short-term memory acts as blackboard; dynamic programming for segmentation; local graph matching for intermediate-level representation; relational database operations; production system for high-level representation; confidence measures for region-object
Outdoor Scenes						
NAOS	B. Neumann H. Novak	University of Hamburg	Neumann and Novak, 1983	Street and traffic scenes	Case frames based on verbs of locomotion hierarchically organized; 3-D shape; temporal knowledge; IS-A; PART-OF	Linear programming for matching; expectations in time; question answering (qv) and connection with a natural-language system
Not available (see Figure 14)	Y. Ohta	Kyoto University	Ohta, 1980	Outdoor color scenes of sky, trees, buildings, and roads	2-D spatial knowledge; color parameter representation; regions and attributes; rules for object properties and relations	Rule-based reasoning; coarse-to-fine region growing; rule applicability ranked on correctness value; focus on best rules for execution

Table 1. *(continued)*

Name	Authors	Institution	References	Domain	Representation	Control
SCHEMA	J. Brolio R. Collins B. Draper A. Hanson E. Riseman	University of Massachusetts at Amherst	Draper, Collins, Brolio, Hanson, and Riseman, 1989	Outdoor scenes of roads and houses	Object schemas; PART-OF; contexts and scenes; 5-point certainty scale; rules; strategies associated with schemas	Blackboard; hierarchical knowledge source organization; expectation; generation; data and goal-driven processing; distributed processing architecture; large number of specialized knowledge sources for computing wide variety of scene physical parameters
VISIONS (see Figure 15)	A. Hanson E. Riseman and many others	University of Massachusetts at Amherst	Hanson and Riseman, 1978; 1984; Parma, Hanson, and Riseman, 1980; York, Hanson, and Riseman, 1981	Outdoor color scenes of houses and trees	Initial development: 2-D spatial knowledge; 3-D spatial knowledge; schemata organized along PART-OF and IS-A; more recent development; rules for object hypothesis and focus of attention	Initial development: blackboard communication; processing cones and relaxation for edge and region extraction; procedural knowledge representation; more recent development: rule-based focus of attention; region and line algorithms without relaxation; intermediate grouping and organizational processes; sensor and representation fusion during interpretation; knowledge-directed feedback to low level processing; some effort to integrate evidential reasoning
Indoor Scenes						
ABLS (Address Block Location System)	S. Srihari C. Wang	State University of New York at Buffalo	Wang and Srihari, 1988	Localization of mail labels	Dependency graph organization for knowledge sources; rules with confidence values; Dempster-Shafer evidence combination; statistical mail database; hypothesis, block and context frames	Three-level hierarchical blackboard; top-down and bottom-up control
Not available	T. Garvey	SRI International	Garvey, 1976	Office scenes of known objects, telephones, desks, and chairs	3-D spatial knowledge; relations; objects as conjunctions of histograms of local features; regions are lists of image samples or bounding polygons in space	Based on planning of operator sequences; plans represented as AND/OR tree; involved three stages, acquire samples, validate and bound to object model; operators are object specific; cost/confidence scoring measures
IGS (see Figure 16)	H. Barrow J. Tenenbaum	SRI International	Tenenbaum and Barrow, 1977	Rooms, mechanical equipment, and landscapes	2-D spatial knowledge; region based; relational constraints; object models as 3-D polyhedral representations	Generalized Waltz filtering; semantic region growing; visibility matrix for 3-D models computed using camera model
PSEIKI	K. Andress A. Kak	Purdue University	Andress and Kak, 1988	Hallway-following and sidewalk-following mobile robot	Maps; production system in OPS3; hierarchy of scenes, objects, faces, edges, vertices; geometric constraints	Blackboard control; OPS3 demons; Dempster-Shafer evidential reasoning in hierarchical space; expectation generation based on current scene beliefs
Medical Images						
ALVEN (see Figure 17)	H. Covvey J. Mylopoulos J. Tsotsos S. Zucker	University of Toronto	Tsotsos and co-workers, 1980; Tsotsos, 1985; 1987	Evaluation of human left ventricular performance from X-ray movie	2-D spatial knowledge; spatiotemporal representation; frames organized with IS-A, PART-OF, similarity temporal precedence; slots have attached interslots constraints for verification and instantiation	Combination of model, goal, data, failure, and temporally directed hypothesis activation; temporal cooperative computation for hypothesis certainty driven by knowledge organization semantics; temporal expectation generation; expectation failure handled by prediction generalization (see Figure 17)
Not available	D. Ballard C. Brown J. Feldman	University of Rochester	Ballard, Brown, and Feldman, 1978	Identification of ribs in chest radiograph	See entry under Aerial Photographs	See entry under Aerial Photographs

Table 1. *(continued)*

Name	Authors	Institution	References	Domain	Representation	Control
Not available	F. Ferrie M. Levine S. Zucker	McGill University	Ferrie, Levine, and Zucker, 1982	Tracking cell motion and morphology in microphotograph image sequences	2-D spatial knowledge; motion knowledge, including shape changes; region based; cell state changes encoded as rules	Views next state prediction and best-match selection as minimization problems; solution similar in form to a Newton-Raphson method; rule interpreter for cell identification and state changes

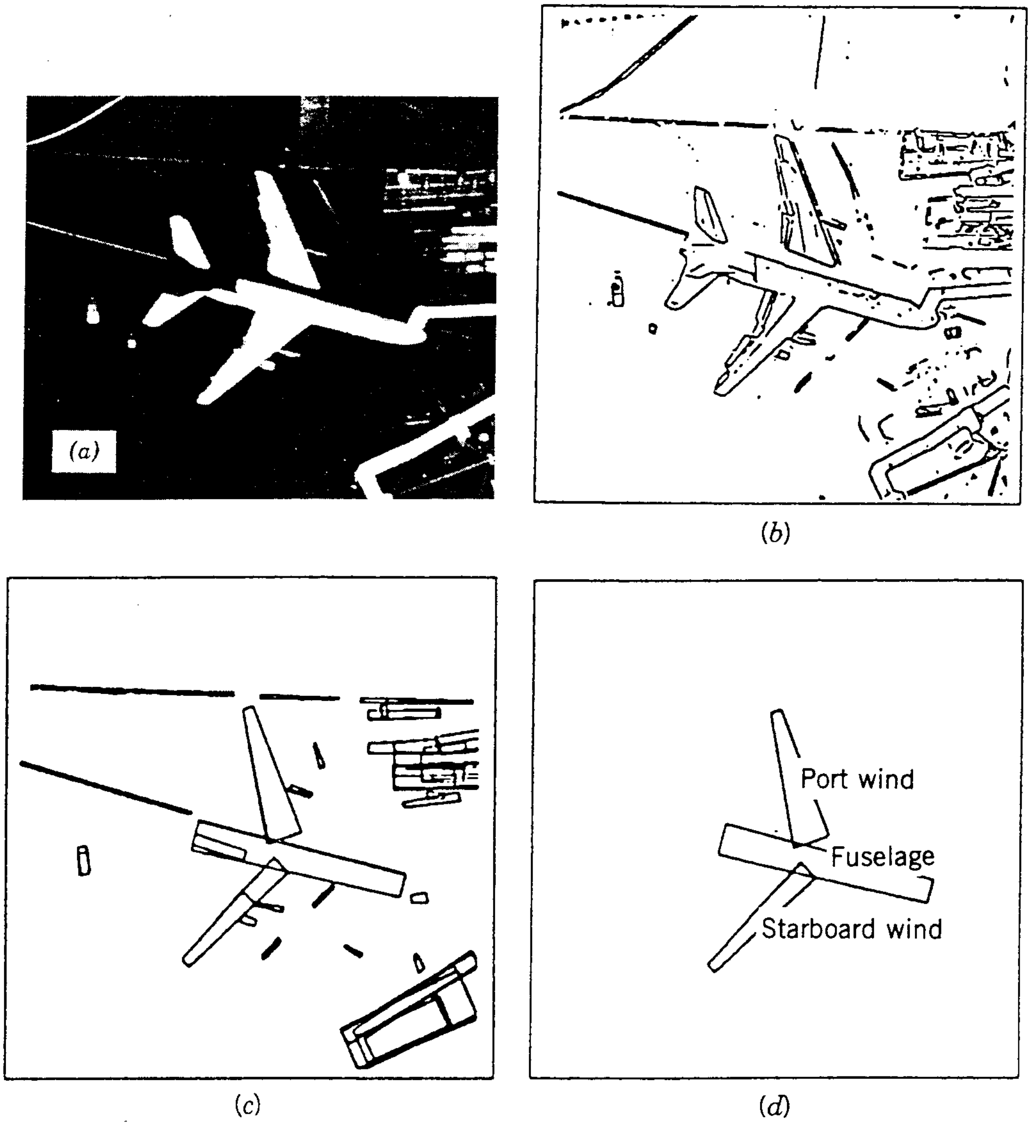
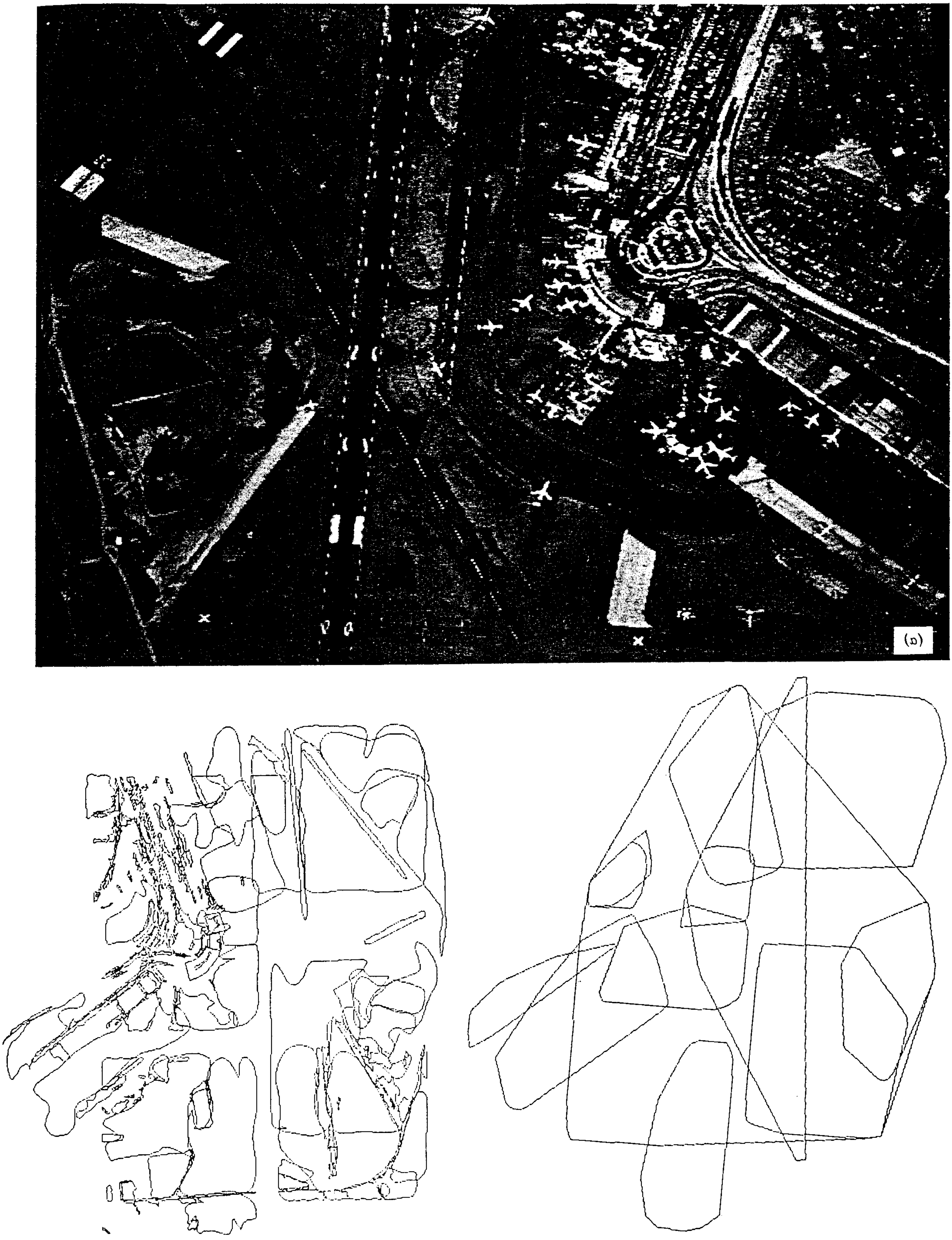


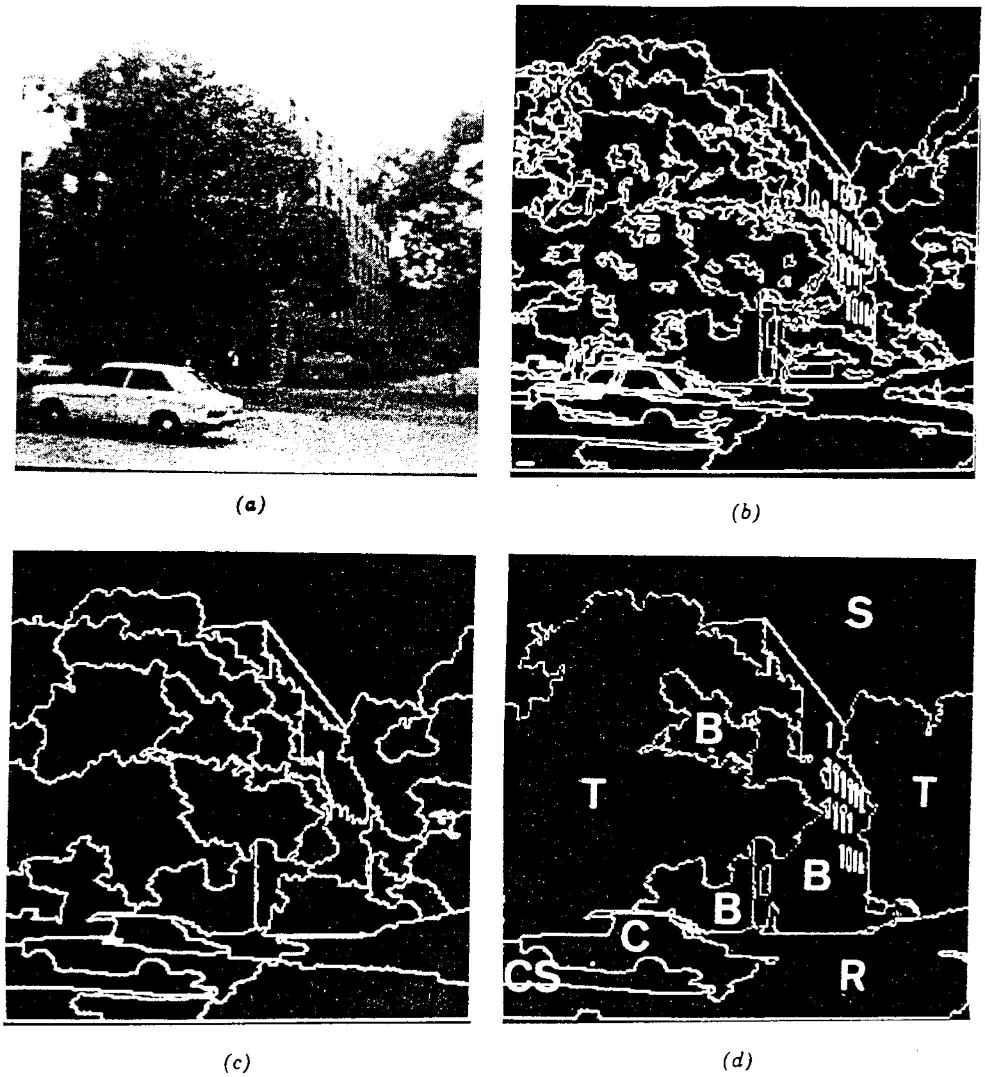
Figure 12. Example of the input and output from the ACRONYM system (Brooks, 1983). An original image is shown, with three steps toward the labeling of the fuselage and wings (*a–d*).

Figure 13. (p. 657) Example of the input and output from the SPAM system (McKeown and co-workers, 1984). (*a*) Original image of an airport scene; (*b*) region-based segmentation produced by SPAM; (*c*) the functional areas extracted by the system.



(b)

(c)


(a)

(b)

(c)

(d)

Figure 14. Example of input and output from Ohta's system (1980). (*a*) digitized input scene; (*b*) result of preliminary segmentation; (*c*) plan image; (*d*) result of meaningful segmentation (S = sky, T = tree, B = building, R = road, C = car, CS = car shadow).

sample input and output of a system is provided. Where more than one example is given, it will be for the purpose of illustrating performance of different control structures. The reader should not assume that the omission of examples for a particular system is a statement on the system's quality.

RESEARCH ISSUES

There are a great many issues outstanding in the field. Perhaps the most important one, and one that is not unique to IU, is the need for a scientific framework within which to design, describe, experiment, and document experiences in IUS building. Few if any attempts at independent verification of claims made are carried out. In other scientific fields, independent duplication of results is a crucial component of the acceptance of a result as a contribution to the field. The lack of much activity in this area may be due to the lack of an overall framework for vision research; the "big picture" within which individual contributions can be placed and interrelated is missing.

Most of the topics covered in this entry require further research, and many issues have already been mentioned. Additional topics specifically addressing the open problems of the IU field are given below.

What Is the Role of Domain Knowledge? Is its application always necessary or does its application depend, perhaps, on the complexity of the scenes being interpreted? Many researchers in the vision community contend that most, if not all, visual interpretation tasks can be carried

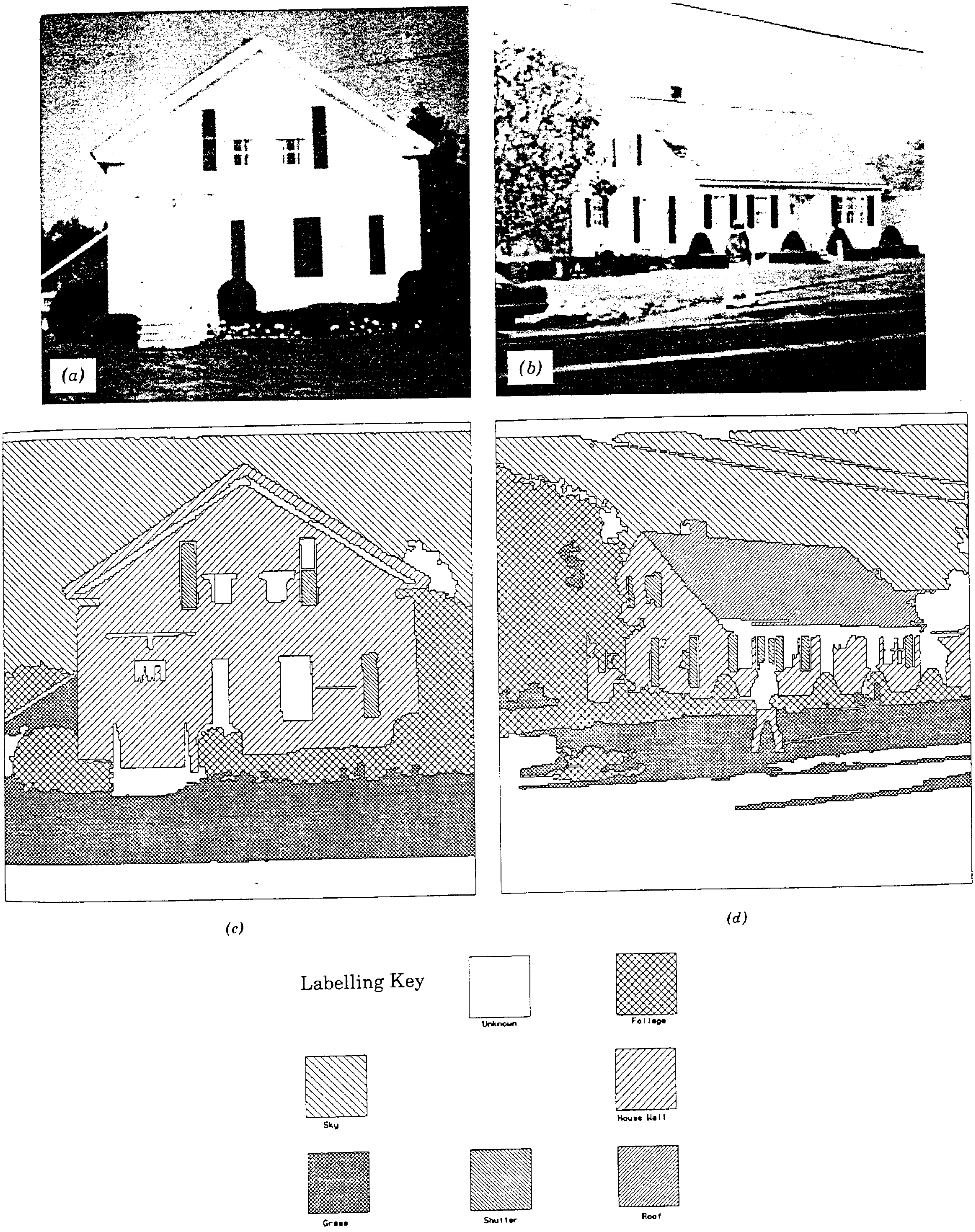
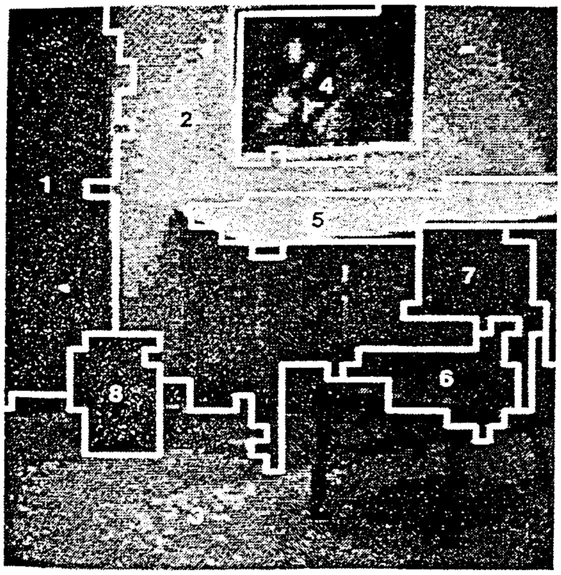


Figure 15. Example of input and output from the VISIONS system (Riseman and Hanson, 1984). (*a*, *b*) Original images; (*c*, *d*) final segmentation and labeling.



Final Region Interpretations

Interpretations	Regions
Door	1
Wall	2
Floor	3
Picture	4
Tabletop	5
Chairseat	6
Chairback	7
Waste Basket	8

Figure 16. Example of input and output from the IGS system (Tenenbaum and Barrow, 1977).

out without domain knowledge (qv), and this issue needs to be explored more fully. A growing segment of the psychology community raises the distinction between attentive and preattentive vision. These are fundamentally different from the high-level/low-level distinctions that computer vision draws and explicitly address the goals of a system in viewing a particular scene as well as scene complexity. The two visual processes are distinguished by their parallel or serial nature, and domain knowledge may play a role in each.

How Can the Best of the Early-Vision Schemes Be Integrated with High-Level Schemes in a Coherent Manner? There currently seems to be no real relationship between the techniques used to extract image features and those used to interpret them. Yet there must be an effective interface, if not also efficient representation transformations, in biological systems.

What Is the Nature of Top-Down Feedback? Does this only impact search schemes, or could it also play a role in expectation generation, in fine tuning of image operators, in priming of semantic concepts, or in bridging the gap between image-centered and world-centered representations, and if so, how? Tsotsos has proven that visual tasks (specifically those that can be cast as visual search problems) become significantly easier if task information is provided (1989) in the same way as model-based vision uses models (see OBJECT RECOGNITION).

Does There Exist a Sufficient Set of Image Features for Image Interpretation?

What Should be Done in Parallel and What Serially; Why and How? How can computations be coordinated and organized?

What is the Nature of the Mechanism That Allows for the Combination of Evidence or Response Strengths?

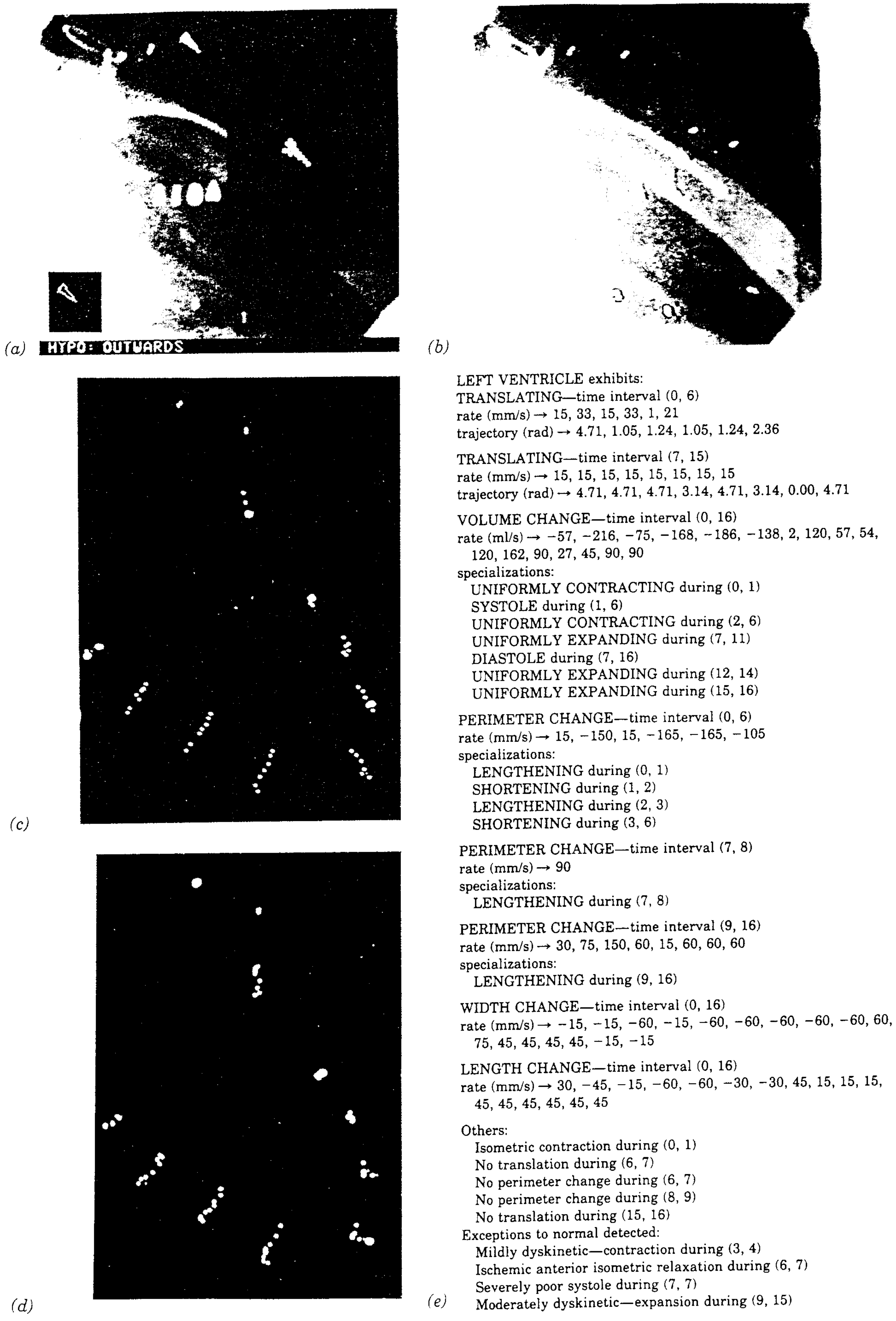
How Can the Biological Sciences Motivate the Design of Image-Understanding Systems? What goes on between the input and the output is a totally unconstrained process, and this points to the major objective in this field: the discovery of computational models that can transform images plus world knowledge into scene interpretations. Guidance from biological research on vision can assist in providing some constraints on the characteristics of the interpretation process. In Tsotsos (1988; 1990), an argument is presented which ties together a great deal of neuroanatomy, neurophysiology, and psychophysics with the thread of complexity satisfaction. An architecture for vision systems, both biological and machine, is derived that satisfies the constraints for human-like visual perception. Much more research along this line is needed however.

Does a Representational Formalism Exist That Spans the Many Required for Vision? Biological vision seems to be nonlinear, time varying, hierarchical, and parallel with a superimposed serial component. Do formalisms exist that can deal with this?

How Can Vision Systems be Integrated into Purposeful, Intelligent Agents Such As Mobile Robots? This research issue incorporates not only topics of vision, sensor and robot control, but also most of AI, specifically knowledge representation, planning, problem-solving, and human-machine communication interfaces.

What Exactly Can Be Learned From System Building? The bottom line is that although image-understanding systems can be engineered to perform reasonably for a tightly constrained domain, the engineering is not yet completely based on sound scientific principles. There is still a long way to go.

Figure 17. Example of input and output from the ALVEN system (Tsotsos, 1985). (*a*) Example of marker finding using motion hypothesis predictions; (*b*) highlighted extracted markers for one image of the image sequence; (*c, d*) inward and outward patterns of motion, respectively, for a complete heart cycle; (*e*) textual output describing the performance characteristics and anomalies detected by ALVEN.



LEFT VENTRICLE exhibits:
TRANSLATING—time interval (0, 6)
rate (mm/s) → 15, 33, 15, 33, 1, 21
trajectory (rad) → 4.71, 1.05, 1.24, 1.05, 1.24, 2.36

TRANSLATING—time interval (7, 15)
rate (mm/s) → 15, 15, 15, 15, 15, 15, 15, 15
trajectory (rad) → 4.71, 4.71, 4.71, 3.14, 4.71, 3.14, 0.00, 4.71

VOLUME CHANGE—time interval (0, 16)
rate (ml/s) → −57, −216, −75, −168, −186, −138, 2, 120, 57, 54, 120, 162, 90, 27, 45, 90, 90
specializations:
UNIFORMLY CONTRACTING during (0, 1)
SYSTOLE during (1, 6)
UNIFORMLY CONTRACTING during (2, 6)
UNIFORMLY EXPANDING during (7, 11)
DIASTOLE during (7, 16)
UNIFORMLY EXPANDING during (12, 14)
UNIFORMLY EXPANDING during (15, 16)

PERIMETER CHANGE—time interval (0, 6)
rate (mm/s) → 15, −150, 15, −165, −165, −105
specializations:
LENGTHENING during (0, 1)
SHORTENING during (1, 2)
LENGTHENING during (2, 3)
SHORTENING during (3, 6)

PERIMETER CHANGE—time interval (7, 8)
rate (mm/s) → 90
specializations:
LENGTHENING during (7, 8)

PERIMETER CHANGE—time interval (9, 16)
rate (mm/s) → 30, 75, 150, 60, 15, 60, 60, 60
specializations:
LENGTHENING during (9, 16)

WIDTH CHANGE—time interval (0, 16)
rate (mm/s) → −15, −15, −60, −15, −60, −60, −60, −60, −60, 60, 75, 45, 45, 45, 45, −15, −15

LENGTH CHANGE—time interval (0, 16)
rate (mm/s) → 30, −45, −15, −60, −60, −30, −30, 45, 15, 15, 15, 45, 45, 45, 45, 45, 45

Others:
Isometric contraction during (0, 1)
No translation during (6, 7)
No perimeter change during (6, 7)
No perimeter change during (8, 9)
No translation during (15, 16)
Exceptions to normal detected:
Mildly dyskinetic—contraction during (3, 4)
Ischemic anterior isometric relaxation during (6, 7)
Severely poor systole during (7, 7)
Moderately dyskinetic—expansion during (9, 15)

BIBLIOGRAPHY

J. Allen, "Towards a General Theory of Action and Time," *Artif. Intell.* **23,** 123–154 (1984).

K. Andress and A. Kak, "Evidence Accumulation and Flow of Control in a Hierarchical Spatial Reasoning System," *AI Magazine* **9**(2), 75–94 (1988).

N. Badler, *Temporal Scene Analysis: Conceptual Descriptions of Object Movements,* Technical Report 80, Department of Computer Science, University of Toronto, 1975.

M. Baird and M. Kelly, "A Paradigm for Semantic Picture Recognition," *Patt. Recog.* **6,** 61–79 (1974).

D. Ballard, C. Brown, and J. Feldman, "An Approach to Knowledge-Directed Image Analysis," in A. Hanson and E. Riseman, eds., *Computer Vision Systems,* Academic Press, New York, 1978, pp. 271–282.

H. Barrow and J. Tenenbaum, "Recovering Intrinsic Scene Characteristics from Images," in A. Hanson and E. Riseman, eds., *Computer Vision Systems,* Academic Press, New York, 1978, pp. 3–26.

H. Barrow, R. Bolles, T. Garvey, T. Kremers, J. Tenenbaum, and H. Wolf, "Experiments in Map-Guided Photo Interpretation," *Proceedings of the Fifth IJCAI,* Cambridge, Mass., 1977, p. 696.

I. Beiderman, "Aspects and Extensions of a Theory of Human Image Understanding," in Z. Pylyshyn, ed., *Computational Processes in Human Vision,* Ablex Press, 1988, pp. 370–428.

T. Binford, "Survey of Model-based Image Analysis Systems," *Int. J. Robot. Res.* 1(1), 18–64 (Spring 1982).

A. Bobick and R. Bolles, "Representation Space: An Approach to the Integration of Visual Information," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* San Diego, Calif., 1989, pp. 492–499.

R. Brooks, "Symbolic Reasoning Among Three-Dimensional Models and Two-Dimensional Images," *Artif. Intell.* **17,** 285–348 (1981).

R. Brooks, "Model-based Three-Dimensional Interpretations of Two-dimensional Images," *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI–5**(2), 140–150 (March 1983).

R. Brooks, R. Greiner, and T. Binford, "The ACRONYM Model-Based Vision System," *Proceedings of the Sixth IJCAI,* Tokyo, Japan, 1979, pp. 105–113.

A. Califano, R. Kjeldsen, and R. Bolle, "Data and Model Driven Foveation," *Proceedings of the International Conference on Pattern Recognition,* Atlantic City, N.J., 1990.

M. Clowes, "On Seeing Things," *Artif. Intell.* **2,** 79–116 (1971).

B. Draper, R. Collins, J. Brolio, A. Hanson, and E. Riseman, "The Schema System," *Int. J. Comput. Vision* **2–3,** 209–250 (1989).

L. Erman, F. Hayes-Roth, V. Lesser, and R. Reddy, "The HEARSAY-II Speech Understanding System: Integrating Knowledge to Resolve Uncertainty," *Comput. Surv.* **12,** 213–253 (1980).

G. Falk, "Interpretation of Imperfect Line Data as a Three-Dimensional Scene," *Artif. Intell.* 3(2), 101–144 (1972).

F. Ferrie, M. Levin, and S. Zucker, "Cell Tracking: A Modeling and Minimization Approach," *IEEE Trans. Pattern Annal. Machine Intell.* **PAMI–4**(3), 277–290 (1982).

M. Fischler and O. Firschein, *Readings in Computer Vision,* Morgan-Kaufmann Press, San Mateo, Calif., 1987.

J. Freeman, "Survey: The Modeling of Spatial Relations," *Comput. Vis. Graph. Img. Proc.* **4,** 156–171 (1975).

E. Freuder, "A Computer System for Visual Recognition Using Active Knowledge," *Proceedings of the Fifth IJCAI,* Cambridge, Mass., 1977, pp. 671–677.

T. Garvey, *Perceptual Strategies for Purposive Vision,* SRI Technical note 117, SRI International, Menlo Park, Calif., 1976.

A. Hanson and E. Riseman, eds., *Computer Vision Systems,* Academic Press, New York, 1978.

A. Hanson and E. Riseman, "VISIONS: A Computer System for Interpreting Scenes," in A. Hanson and E. Riseman, eds., *Computer Vision Systems,* Academic Press, New York, 1978, pp. 303–334.

A. Huertas, W. Cole, and R. Nevatia, "Using Generic Knowledge in Analysis of Aerial Scenes; A Case Study," *Proceedings of the Eleventh IJCAI,* Detroit Mich., Morgan-Kaufmann, San Mateo, Calif., 1989, pp. 1642–1648.

D. Huffman, "Impossible Objects as Nonsense Sentences," *Artif. Intell.* **3,** 295–323 (1971).

R. Hummel and S. Zucker, *On the Foundations of Relaxation Labelling Processes,* Technical Report 80–7, Department of Electrical Engineering, McGill University, Montreal, 1980.

A. Jepson and W. Richards, "What is a Percept?" *Cogn. Sci.* (submitted, 1990).

T. Kanade, "Model Representations and Control Structures in Image Understanding," *Proceedings of the Fifth IJCAI,* Cambridge, Mass., Morgan-Kaufmann, San Mateo, Calif., 1977, pp. 1074–1082.

T. Kanade, "Survey: Region Segmentation: Signal vs Semantics," *Comput. Vis. Graph. Img. Process.* **13,** 279–297 (1980).

M. Kass, A. Witkin, and D. Terzopoulos, "SNAKES: Active Contour Models," *Int. J. Comput. Vision* **1-4,** 321–332 (1988).

M. Kelly, "Edge Detection in Pictures by Computer Using Planning," *Machine Intell.* **6,** 397–409 (1971).

B. Kuipers, "Modelling Spatial Knowledge," *Cogn. Sci.* **2,** 129–154 (1978).

M. Levine, "A Knowledge-Based Computer Vision System," in A. Hanson and E. Riseman, eds., *Computer Vision Systems,* Academic Press, New York, 1978, pp. 335–352.

B. Lowerre and R. Reddy, "The HARPY Speech Understanding System," in W. A. Lea, ed., *Trends in Speech Recognition,* Prentice-Hall, Englewood Cliffs, N.J., 1980, Chapter 15.

D. McDermott and E. Davis, "Planning Routes Through Uncertain Territory," *Artif. Intell.* **22,** 107–156 (1984).

D. McKeown, W. Harvey, and J. McDermott, "Rule-Based Interpretation of Aerial Imagery," *Proceedings of the IEEE Workshop on Principles of Knowledge-Based Systems,* Denver, Colo., 1984, pp. 145–158.

A. Mackworth, "On Reading Sketch Maps," *Proceedings of the Fifth IJCAI,* Cambridge, Mass., Morgan-Kaufmann, San Mateo, Calif., 1977, pp. 598–606.

A. Mackworth, "Vision Research Strategy: Black Magic, Metaphors, Mechanisms, Miniworlds, and Maps," in A. Hanson and E. Riseman, eds., *Computer Vision Systems,* Academic Press, New York, 1978, pp. 53–60.

A. Mackworth and W. Havens, "Representing Knowledge of the Visual World," *IEEE Comput.* **16,** 90–98 (1983).

D. Marr, *Vision,* W. H. Freeman, San Francisco, Calif., 1982.

D. Marr and H. Nishihara, "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes," *Proc. R. Soc. London Ser. B* **200,** 269–294 (1978).

T. Matsuyama, "Knowledge Organization and Control Structure in Image Understanding," *Proceedings of the ICPR,* Montreal, Quebec, 1984, pp. 1118–1127.

T. Matsuyama and V. Hwang, "SIGMA: A Framework for Image

Understanding: Integration of Bottom-Up and Top-Down Analyses," *Proceedings of the Ninth IJCAI,* Los Angeles, Calif., Morgan-Kaufmann, San Mateo, Calif., 1985, pp. 908–915.

M. Minsky, "A Framework for Representing Knowledge," in P. Winston, ed., *The Psychology of Computer Vision,* McGraw-Hill, New York, 1975, pp. 211–277.

J. Mulder, A. Mackworth, and W. Havens, "Knowledge Structuring and Constraint Satisfaction: The Mapsee Approach," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-10-6,** 866–879 (1988).

M. Nagao and T. Matsuyama, *A Structural Analysis of Complex Aerial Photographs,* Plenum, New York, 1980.

B. Neumann and H. Novak, "Event Models for Recognition and Natural Language Description of Events in Real-World Image Sequences," *Proceedings of the Eighth IJCAI,* Karlsruhe, FRG, Morgan-Kaufmann, San Mateo, Calif., 1983, pp. 724–726.

R. Nevatia, "Characterization and Requirements of Computer Vision Systems," in A. Hanson and E. Riseman, eds., *Computer Vision Systems,* Academic Press, New York, 1978, pp. 81–88.

Y. Ohta, "A Region-Oriented Image Analysis System by Computer," Ph.D. dissertation, Kyoto University, Department of Information Science, 1980.

J. O'Rourke and N. Badler, "Model-based Image Analysis of Human Motion Using Constraint Propagation," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-2,** 522–536 (1980).

C. Parma, A. Hanson, and E. Riseman, "Experiments in Schema-Driven Interpretation of a Natural Scene," *COINS TR* 80–10, University of Massachusetts at Amherst, 1980.

A. Pentland, "Perceptual Organization and the Representation of Natural Form," *SRI Technical Note* 357, SRI International, Menlo Park, Calif., 1985.

E. Riseman and A. Hanson, "A Methodology for the Development of General Knowledge-Based Vision Systems, *Proceedings of the IEEE Workshop on Principles of Knowledge-Based Systems,* Denver, Colo. 1984, pp. 159–172.

L. Roberts, "Machine Perception of Three-Dimensional Solids," in J. Tippett and co-workers, eds., *Optical and Electro-optical Information Processing,* MIT Press, Cambridge, Mass., 1965, pp. 159–197.

S. Rubin, "Natural Scene Recognition Using LOCUS Search," *Comput. Vis. Graph. Img. Process.* **13,** 298–333 (1980).

Y. Shirai, "A Context-sensitive Line Finder for Recognition of Polyhedra," *Artif. Intell.* 4(2), 95–119 (1973).

S. Tanimoto and A. Klinger, eds., *Structured Computer Vision,* Academic Press, New York, 1980.

J. Tenenbaum and H. Barrow, "Experiments in Interpretation Guided Segmentation," *Artif. Intell.* **8**(3), 241–274 (1977).

D. Terzopoulos, A. Witkin, and M. Kass, "Constraints on Deformable Models: Recovering 3-D Shape and Nonrigid Motion," *Artif. Intell.* **36**(1), 91–123 (1988).

J. Tsotsos, "Knowledge of the Visual Process: Content, Form and Use," *Patt. Recog.* **17,** 13–28 (1984).

J. Tsotsos, "Knowledge Organization and Its Role in the Interpretation of Time-varying Data: The ALVEN System," *Computat. Intell.* 1(1), 16–32 (1985).

J. Tsotsos, "Representational Axes and Temporal Cooperative Computation," in M. Arbib and A. Hanson, eds., *Vision, Brain and Cooperative Computation,* MIT Press, Bradford Books, Cambridge, Mass., 1987, pp. 361–418.

J. Tsotsos, "A Complexity Level Analysis of Immediate Vision," *Int. J. Comput. Vision* **1-4,** 303–320 (1988).

J. Tsotsos, "The Complexity of Perceptual Search Tasks," Proceedings of the Eleventh IJCAI, Detroit, Mich., Morgan-Kaufmann, San Mateo, Calif., 1989.

J. Tsotsos, "A Complexity Level Analysis of Vision," *Behavioral and Brain Sciences,* **12**(3) (1990).

J. Tsotsos, J. Mylopoulos, H. Covvey, and S. Zucker, "A Framework for Visual Motion Understanding," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-2,** 563–573 (1980).

L. Uhr, "Layered 'Recognition Cone' Networks that Preprocess, Classify, and Describe," *IEEE Trans. Comput.* **21,** 758–768 (1972).

W. Uttal, *A Taxonomy of Visual Processes,* Lawrence Erlbaum, Hillsdale, N.J., 1981.

C. Wang and S. Srihari, "A Framework for Object Recognition in a Visually Complex Environment and its Application to Locating Address Blocks on Mail Pieces," *Int. J. Comput. Vision* **2-2,** 125–152 (1988).

W. Woods, "Theory Formation and Control in a Speech Understanding System with Extrapolations towards Vision," in A. Hanson and E. Riseman, eds., *Computer Vision Systems,* Academic Press, New York, 1978, pp. 379–380.

B. York, A. Hanson, and E. Riseman, "3-D Object Representations and Matching with B-Splines and Surface Patches," *Proceedings of the Seventh IJCAI,* Vancouver, B.C., Canada, Morgan-Kaufmann, San Mateo, Calif., 1981, pp. 648–651.