NBSIR 82-2582

AN OVERVIEW OF COMPUTER VISION

September 1982

Prepared for

National Aeronautics and Space Administration Headquarters
Washington, D.C. 20546

NBSIR 82-2582

AN OVERVIEW OF COMPUTER VISION

William B. Gevarter*

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
National Engineering Laboratory
Center for Manufacturing Engineering
Industrial Systems Division
Metrology Building, Room A127
Washington, DC 20234

September 1982

Prepared for:
National Aeronautics and Space Administration
Headquarters
Washington, DC 20546

U.S. DEPARTMENT OF COMMERCE, Malcolm Baldrige, *Secretary*
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, *Director*

*Research Associate at the National Bureau of Standards Sponsored by NASA Headquarters

Preface

Computer Vision*

Computer Vision -- visual perception employing computers -- shares with "Expert Systems" the role of being one of the most popular topics in Artificial Intelligence today. Commercial vision systems have already begun to be used in manufacturing and robotic systems for inspection and guidance tasks. Other systems at various stages of development, are beginning to be employed in military, cartograhic and image interpretation applications.

This report reviews the basic approaches to such systems, the techniques utilized, applications, the current existing systems, the state-of-the-art of the technology, issues and research requirements, who is doing it and who is funding it, and finally, future trends and expectations.

The computer vision field is multifaceted, having many participants with diverse viewpoints, with many papers having been written. However, the field is still in the early stages of development--organizing principles have not yet crystalized, and the associated technology has not yet been rationalized. Thus, this report is not as smooth and even as would be desirable. Nevertheless, this overview should prove useful to engineering and research managers, potential users and others who will be affected by this field as it unfolds.

*This report is in support of the more general NBS/NASA report, An Overview of Artificial Intelligence and Robotics.

Acknowledgments

I wish to thank the many individuals and organizations who have contributed to this report by furnishing information and suggestions. I particularly would like to thank Drs. T. Binford of Stanford, M. Brady of M.I.T., M. Tenenbaum and H. Barrow of Fairchild and Gennery et al of JPL for furnishing source material that was essential to the development of this report. In addition, I would like to thank the staff of the NBS Industrial Systems Division, CDMR R. Ohlander of DARPA, Drs. A. Rosenfeld and M. Schneer of U. of MD., M. Fischler of SRI, E. Sacerdoti of M.I.C. and J. Wilder of Object Recognition Systems for reviewing this report and suggesting corrections and modifications. However, any remaining errors or omissions must remain the responsibility of the author. I would also like to thank Margie Johnson for typing and facilitating the publication of this NBS series of reports on Artificial Intelligence and Robotics.

It is not the intent of the National Bureau of Standards to recommend or endorse any of the manufacturers or organizations named in this report, but simply to attempt to provide an overview of the computer vision field. However, in such a diverse and rapidly changing field, important activities and products may not have been mentioned. Lack of such mention does not in any way imply that they are not also worthwhile. The author would appreciate having any such omissions or oversights called to his attention so that they can be considered for future reports.

Table of Contents

Figures

Tables

Computer Vision

I. Introduction

Following the lead of Cohen and Feigenbaum (1982, p. 127) we may consider computer vision to be the information-processing task of understanding a scene from its projected images. Other fields such as image processing and pattern recognition also utilize computers in vision tasks. However, we can distinguish the fields by categorizing them as follows:

Image processing is a signal processing task that transforms an input image into a more desirable output image through processes such as noise reduction, contrast enhancement and registration.

Pattern recognition is a classification task that classifies images into predetermined categories.

Computer vision is an image understanding task that automatically builds a description not only of the image itself, but of the three dimensional scene that it depicts. The term scene analysis has been used in the past to emphasize the distinction between processing two dimensional images, as in pattern classificiation, and seeking information about three-dimensional scenes.

In this report, we will emphasize the Artificial Intelligence (AI) aspects of vision and therefore will dwell on image understanding. Image understanding includes among its techniques, many of the methods found in image processing and pattern recognition. However, it also includes geometric

modeling, and AI knowledge representation and cognitive processing techniques.

Hiatt (1981, p. 3) observes, "For practical purposes, investigators of computer vision often define seeing as gathering visual data for the purpose of making complex decisions. Computer vision is accordingly, a major adjunct to the study of artificial intelligence." Arden (1980, p. 482) adds, "A view widely held by psychologists is that perception is an active process in which hypotheses are formed about the nature of the environment and sensory information is sought that will confirm or refute these hypotheses. This view of perception, as a form of problem-solving at least at some stage, is held by many researchers in artificial intelligence." Thus, computer vision with its many current and potential applications is a major Artificial Intelligence (AI) topic today. The following chapters are an attempt to provide an overview of this important and growing field. In addition to reviewing the conceptual basis for computer vision and its associated techniques, we will also review their implementation in vision systems, both research and commercial.

Chapters II, III and IV further define computer vision, reviewing its origins and its relation to human vision. Chapter V briefly indicates applications of computer vision.

Chapters VI outlines a basis for a general purpose computer vision system, in the process providing a structure for comprehending systems with lesser aspirations.

Chapter VII reviews the basic control structures suitable for vision systems. Chapter VIII examines the successive levels of representation found in computer vision systems.

Vision systems, both research and industrial are covered in Chapters IX and X. Information on who the principal participants are in the computer vision field is given in Chapters XI and XII.

The state-of-the-art, current problems and issues, research requirements and future trends are presented in Chapters XIII to XVI.

Reviews of the representation methods and processing techniques used in computer vision are given in the appendices.

Appendix A reviews representations for low level image features such as pixels, edges, regions, etc.

Appendix B reviews techniques (such as filtering and thresholding) for extracting edges and regions.

Appendix C discusses methods for symbiotically combining image segmentation with interpretation.

Appendix D provides an overview of methods (such as statistical features, boundary curves, primitive forms, and relational graphs) for succinctly representing image features and utilizing the resulting representation for recognition.

Appendix E reviews the various methods for extracting intrinsic image characteristics such as surface shapes, ranges and orientations from 2-D images. Also included is a discussion of extracting shape and velocity from successive images of objects in motion.

Appendix F provides an overview of higher levels of

representation--both volumetric and procedural models, and symbolic descriptions such as relational graphs.

Appendix G reviews how intrinsic images can be given higher level interpretations by segmenting intrinsic surface characteristics into objects (either by model or symbolic description matching) yielding object recognitions or scene descriptions.

Appendix H reviews real-time visual tracking, needed for guidance, assembly and other tasks.

A glossary of terms in computer vision is given in Appendix K. Publications sources for further information are listed in Appendix L.

II. Definition

Computer (computational or machine) vision can be defined as perception by a computer based on visual sensory input.

Horn (1979, pp. 70-71) characterizes machine vision (from a robotic orientation) as follows:

> An optical system forms an image of some three-dimensional [3-D] arrangement of parts. The two-dimensional [2-D] image is sensed and converted into machine readable format. It is the purpose of the machine vision system to derive information from this image useful in the execution of the given task. In the simplest case the information sought will concern only the location and orientation of an isolated object--more commonly, objects have to be recognized and their spatial relationships determined. This can be viewed as a process in which a description of the scene being viewed is developed from the raw image. The description has to be appropriate to the particular application. That is, irrelevant visual features should be discarded, while needed relationships between parts of objects must be deduced from their optical projection.

Barrow and Tenenbaum (1981, p. 573) enlarge on this from a more general viewpoint, stating:

> Vision is an information-processing task with well-defined input and output. The input consists of arrays of brightness values, representing projections of a three-dimensional scene recorded by a camera or comparable imaging device. Several input arrays may provide information in several spectral bands (color) or from multiple viewpoints (stereo or time sequence). The desired output is a concise description of the three-dimensional scene depicted in the image, the exact nature of which depends upon the goals and expectations of the observer. It generally involves a description of objects and their interrelationships, but may also include such information as the three-dimensional structure of surfaces, their physical characteristics (shape, texture, color, material), and the locations of shadows and light sources...

In this report, we will follow the lead of Ballard and Brown (1982, p. 2) and define Computer Vision as "the enterprise of automating and integrating a wide range of processes and

representations used for vision perception." The emphasis will be on generating a description or an understanding of the scene from which the image was obtained. The next chapter will enlarge on this point of view.

III. Origins of Computer Vision

Computer vision is based largely on ideas from three related fields: image processing, pattern recognition and scene analysis.

Rosenfeld (1981, p. 596) states that, "In image processing, the input and the output are both images with the output an improved version of the input." In preprocessing we have gray-scale modification (usually to normalize scene brightness and contrast), sharpening (to restore the weakened high spatial frequencies) and smoothing to remove noise in the image. If two images have to be compared, they may have to be registered (i.e., geometrically transformed to make then congruent) before matching them.

In pattern recognition the input is the image, but the output is a description of the image based on a priori knowledge of expected patterns. The computer usually starts with a list of brightness values associated with the array of hundreds of thousands of points corresponding to the image. Recognizing a pattern means replacing this mass of undigested data with a much simpler more useful description. However, it is usually impractical to search directly for examples of the patterns we are interested in this array of intensity values. Instead, it is often more convenient to first search for examples of simpler patterns (such as edges and regions), referred to as features. A

simplified description of the image constructed from these features can then be used as the basis for pattern recognition*.

Scene analysis is concerned with the transformation of simple features into abstract descriptions relating to objects that cannot be simply recognized based on pattern matching.

Brady (1981A, pp. 4-5) referring to scene analysis as image understanding (IU) expands on the differences between pattern recognition and IU, observing that typically pattern recognition systems are concerned with recognizing the input as one of a usually small set of possibilities. Pattern recognition systems are mostly concerned with images of basically two dimensional objects. When the images are of three dimensional objects, such as engine parts, they are effectively treated as two dimensional, by considering each stable position as a separate object. In contrast, IU has dealt extensively with three dimensional images.

More significantly, pattern recognition systems typically operate directly on the image. IU approaches to most visual processes (e.g., stereo, texture, shape from shading), operate not on the image but on symbolic representations that have been computed by earlier processing such as edge detection.

Arden (1980, pp. 482-483), taking a historical perspective, contrasts the pattern-recognition and the IU or AI approach as follows:

---

* Pratt (1978, pp. 568-569) indicates that in many cases for simple objects in uncluttered imagery, it is feasible to extract needed features by transformations of the images (e.g., using a two dimensional Fourier transform). The resulting feature space can be partioned into regions for classification into objects, based on prototypes.

> Since the early sixties there has been a marked divergence between the pattern-recognition and AI approaches to computer analysis of images. The former approach has continued to stress the use of ad hoc image features in combination with statistical classification techniques. More recently, use has been made of "syntactic" methods in which images are recognized by a "parsing" process as being built up hierarchically of primitive constituents. By contrast, the AI approach has employed problem-solving methodologies based on extensive use of knowledge about the class of images, or "scenes," to be analyzed...
>
> Much of the work on computer vision has dealt with images of scenes containing solid objects viewed from nearby. These are the sort of images with which a robot vision system must cope in using vision to guide its motor activities, including manipulation and locomotion. The analysis of such images is usually called "scene analysis," to distinguish it from the analysis of images that are essentially two-dimensional, such as photomicrographs (which show cross-sections), radiographs (which show projections), satellite imagery (in which terrain relief is negligible), documents, diagrams, maps, and so on. The methods of computer vision, however, apply equally to these latter classes of images; the term need not be restricted to three-dimensional scene analysis.

In this report we will only treat image processing and low-level vision to the extent needed for image understanding. Pattern recognition, which has broken off from AI and has also become a separate field, will also be given minimum treatment.

To a large extent, the terms scene analysis, image understanding, and computer vision have become synonymous. The more advanced vision systems have a strong AI flavor, being heavily concerned with symbolic processes for representing and manipulating knowledge in a problem solving mode. Though vision systems that primarily depend on pattern recognition techniques are also treated in this report, the intent is to concentrate on the knowledge-based scene analysis (IU) approach which is the major focus in AI computational vision.

In the next chapter we will briefly look at the relation that human vision has to the AI approach to computer vision.

IV. Relation to Human Vision

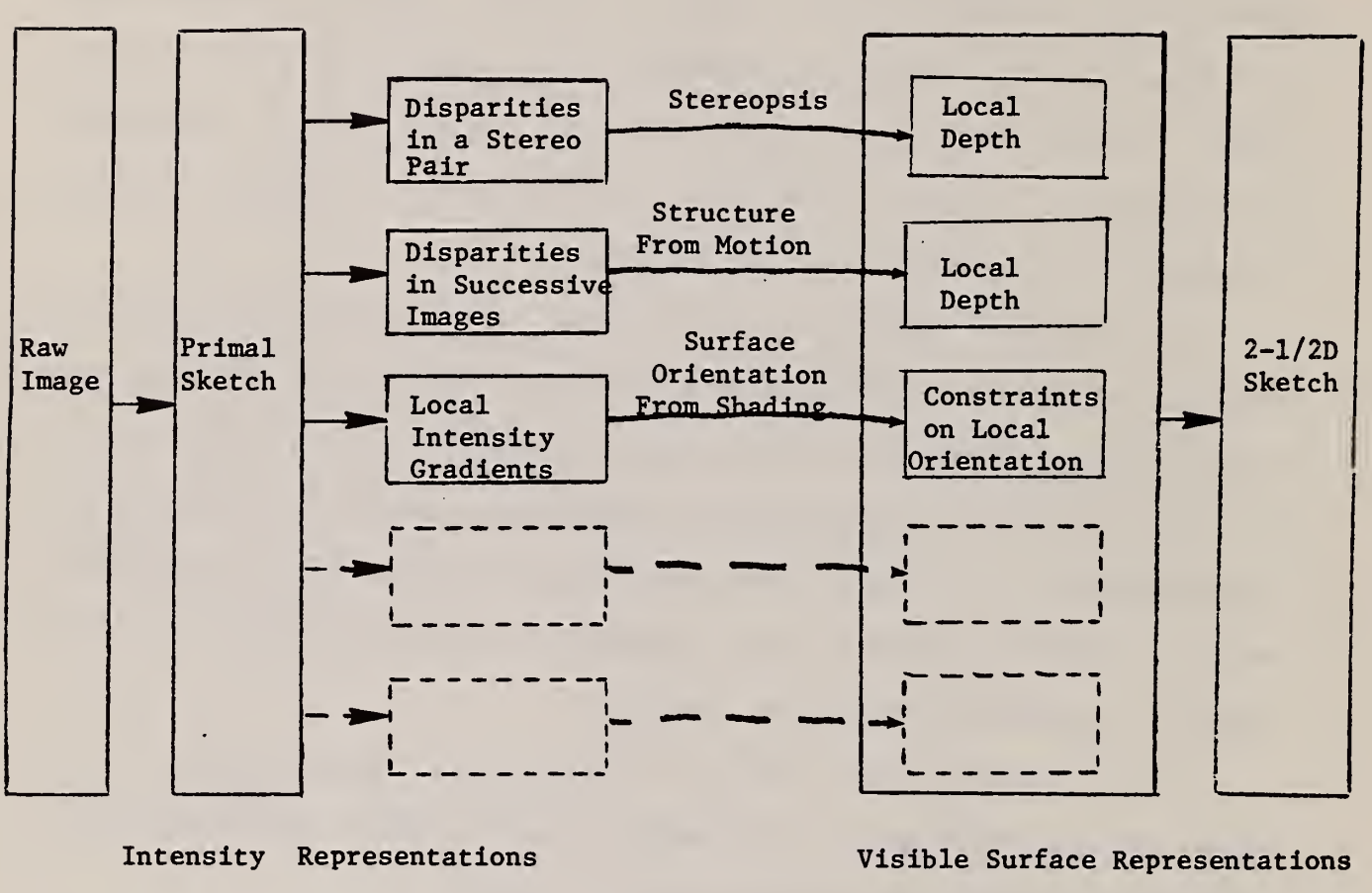
MIT's Marr and Nishihara (1978, p. 42) take the view that "Artificial Intelligence is (or ought to be) the study of information processing problems that characteristically have their roots in some aspect of biological information processing." They developed a computational theory of vision based on their study of human vision. Figure 1 represents the transition from the raw image through the primal sketch to the 2-1/2D sketch (indicated in Figure 2), which contains information on local surface orientations, boundaries, and depths.

The primal sketch, reminiscent of an artist's hurried drawing, is a primitive but rich description of the way the intensities change over the visual field. It can be represented by a set of short line segments separating regions of different brightnesses. A list of the properties of the lines segments, such as location, length, and orientation for each segment can be used to represent the primal sketch.

The late Dr. Marr and his associates' development of a human visual information processing theory (Marr, 1982) has had a substantial impact on computational vision.

Figure 1

A Framework for Early and Intermediate States in
A Theory of Visual Information Processing

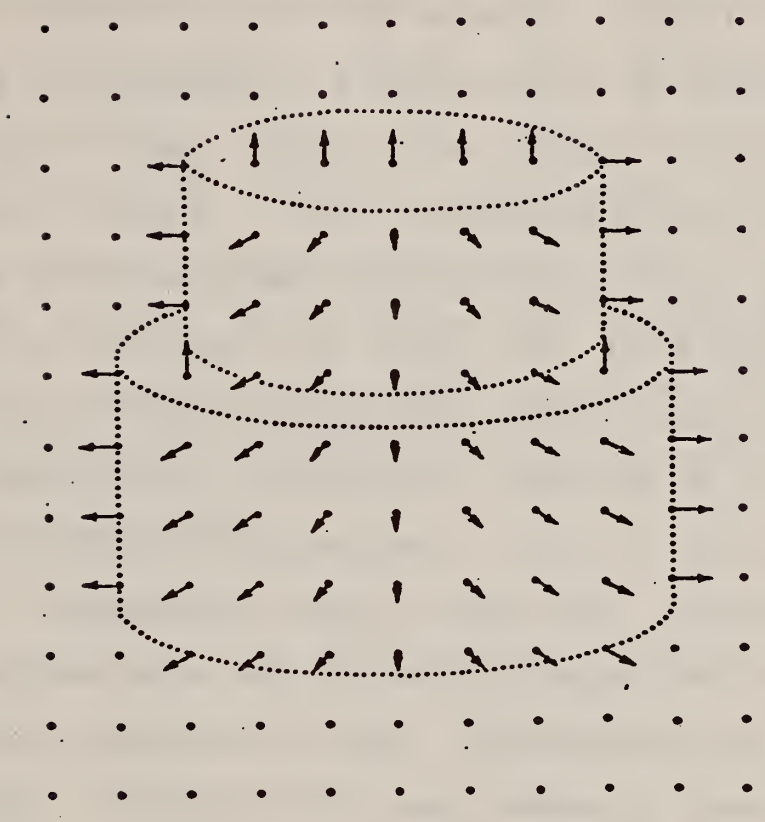


The computations begin with representations of the intensities in an image--first the image itself, (e.g., the gray-level intensity array) and then the primal sketch, a representation of spatial variations in intensity. Next comes the operation of a set of modules, each employing certain aspects of the information contained in the image to derive information about local orientation, local depth, and the boundaries of surfaces. From this is constructed the so-called 2-1/2 dimensional sketch. Note that no "high-level" information is yet brough to bear: the computations proceed by utilizing only what is available in the image itself.

Source: Marr and Nishihara, 1978, p. 42.

Figure 2

An Example of a 2-1/2D Sketch



A candidate for the so-called 2-1/2-dimensional sketch, which encompasses local determinations of the depth and orientation of surfaces in an image, as derived from processes that operate upon the primal sketch or some other representation of changes in gray-level intensity. The lengths of the needles represent the degree of tilt at various points in the surface; the orientations of the needles represent the directions of tilt... Dotted lines show contours of surface discontinuity. No explicit representation of depth appears in this figure.

Source: Marr and Nishihara, 1978, p. 41.

Barrow and Tenenbaum (1981, pp. 579-580) also seek insights into the organization of a high-performance, general-purpose visual system from observations of the behavior of the human visual system. They observe that a person looking at a natural scene, such as the landscape, is aware of many intermediate levels of description, such as surfaces, volumes, and shadows. Over a wide range of viewpoint and illumination, a person can readily estimate quite accurately such local surface characteristics as reflectance, color, texture, distance, and orientation, as well as such global characteristics as size, and shape. Boundaries are seen not merely as intensity discontinuities, but as physically significant events--discontinuities in distance, orientation, reflectance, incident illumination, and so forth. Humans also experience immediate global perceptions: the type of scene (landscape), the dominant orientations of the support plane and the gravitational vertical, the direction of illumination, and the viewpoint with respect to these. Thus, what a person sees are intrinsic characteristics of three-dimensional surfaces, not transient features of a two-dimensional image as observed under a particular set of viewing conditions.

They also note that perception by humans of surfaces and surface boundaries does not appear to depend critically upon contrast nor familiarity with the specific objects depicted.

There are strong indications (c.f. Gevarter, 1977) that the interpretative planning areas of the human brain set up a context for processing the input data. (This is captured by Minsky's (1975) AI "frame" concept for knowledge representation). The

brain then uses visual and other cues from the environment to draw in past knowledge to generate an internal representation and interpretation of the scene. This knowledge-based expectation-guided approach to vision is now appearing in the advanced AI computer vision systems (discussed in later Chapters).

Barrow and Tennenbaum suggest that insights gained by studying human vision, coupled with experience resulting from building machine vision systems, can provide the basis for a computational model of visual processing. Their approach to a general purpose computer vision system will be pursued in Chapter VI, but now we pause to motivate this pursuit by briefly reviewing applications of computer vision already underway in this rapidly growing field.

V. Applications

Brady (1981A, p. 2) states that, "There is currently a surge of interest in image understanding on the part of industry and the military." Current computer vision applications, primarily taken from Brady (1981A, pp. 3-4), are listed in Figure 3.

Figure 3: Examples of Applications of Computer Vision Now Underway

- AUTOMATION OF INDUSTRIAL PROCESSES

  Object acquisition by robot arms, for example sorting or packing items arriving on conveyor belts.

  Automatic guidance of seam welders and cutting tools.

  VLSI-related processes, such as lead bonding, chip alignment and packaging.

  Monitoring, filtering, and thereby containing the flood of data from oil drill sites or from seismographs.

  Providing visual feedback for automatic assembly and repair.

- INSPECTION TASKS

  The inspection of printed ciruit boards for spurs, shorts, and bad connections.

  Checking the results of casting processes for impurities and fractures.

  Screening medical images such as chromosome slides, cancer smears, x-ray and ultrasound images, tomography.

  Routine screening of plant samples.

  Inspection of alpha-numerics on labels and manufactured items.

  Checking packaging and contents in pharmaceutical and food industries.

  Inspection of glass items for cracks, bubbles, etc.

- REMOTE SENSING

  Cartography: the automatic generation of hill-shaded maps, and the registration of satellite images with terrain maps.

  Monitoring traffic along roads, docks, and at airfields.

  Management of land resources such as water, forestry, soil erosion, and crop growth.

  Exploration of remote or hostile regions for fossil fuels and mineral ore deposits.

Figure 3 (cont.)

- MAKING COMPUTER POWER MORE ACCESSIBLE

  Management information systems that have a communication channel considerably wider than current systems that are addressed by typing or pointing.

  Document readers (for those who still use paper).

  Design aids for architects and mechanical engineers.

- MILITARY APPLICATIONS

  Tracking moving objects.

  Automatic navigation based on passive sensing.

  Target acquisition and range finding.

- AIDS FOR THE PARTIALLY SIGHTED

  Systems that read a document and speak what they read.

  Automatic "guide dog" navigation systems.

VI. Basis for a General Purpose Image Understanding System

Barrow and Tenenbaum (1981, p. 573) observe that in going from a scene to an image (an array of brightness values) that the image encodes much information about the scene, but the information is confounded in the single brightness value at each point. In projecting onto the two-dimensional image, information about the three-dimensional structure of the scene is lost. In order to decode brightness values and recover a scene description, it is necessary to employ a priori knowledge embodied in models of the scene domain, the illumination, and the imaging process.

Scene models can be devised to describe the three-dimensional world in terms of surfaces and objects.

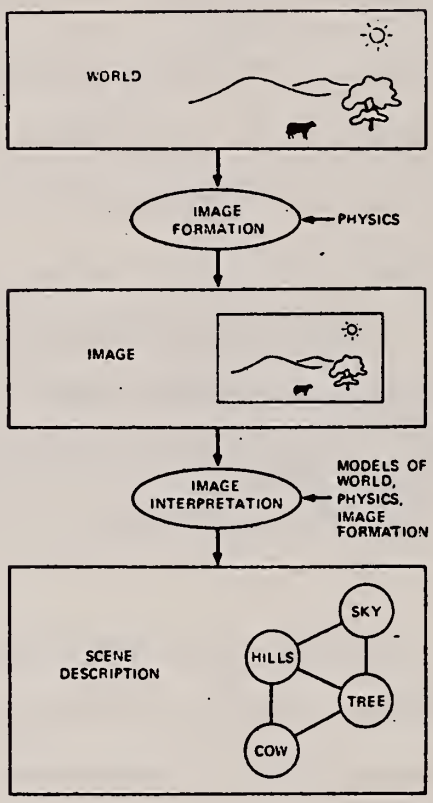
Illumination models can be utilized to describe the primary light sources, their positions, spatial extents, intensities, colors, and so forth.

Sensor models describe the photometric and geometric properties of the sensor, which can be used to predict how a particular scene, observed from a particular viewpoint and under particular illumination conditions, is transformed into the two-dimensional array of brightness values that constitutes the input.

As indicated by Figure 4, computer vision is an active process that uses these models to interpret the sensory data. To accommodate the diversity of appearance found in real imagery, a high-performance, general-purpose system must embody a great deal of knowledge in its models.

Figure 4

Model-based Interpretation of Images



Source: Barrow and Tenenbaum, 1981, p. 573.

The next three chapters review the work in devising computer vision systems. Chapter VII discusses paradigms for computer vision systems. Chapter VIII presents the levels of representation appropriate to high performance systems. Chapter IX reviews research efforts in building such systems.

VII. Basic Paradigms for Computer Vision*

In broad terms, an image understanding system starts with the array of pixel amplitudes that define the computer image, and using stored models (either specific or generic) determines the content of a scene. Typically, various symbolic features such as lines and areas are first determined from the image. These are then compared with similar features associated with stored models to find a match, when specific objects are being sought. In more, generic cases, it is necessary to determine various characteristics of the scene, and using generic models determine from geometric shapes and other factors (such as allowable relationships between objects) the nature of the scene content.

A variety of paradigms have been proposed to accomplish these tasks in image understanding systems. These paradigms are based on a common set of broadly defined processing and manipulating elements: feature extraction, symbolic representation, and semantic interpretation. The paradigms differ primarily in how these elements (defined below) are organized and controlled, and the degree of artificial intelligence and knowledge employed.

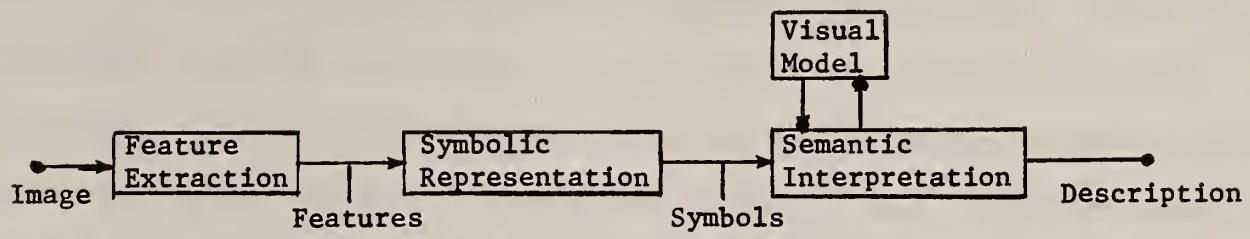
A. Hierarchical-Bottom-up Approach

Figure 5A is a block diagram of a hierarchical paradigm of an image understanding system that employs a bottom-up processing approach. First, primitive features are extracted from the array of picture element intensities that constitute the observed

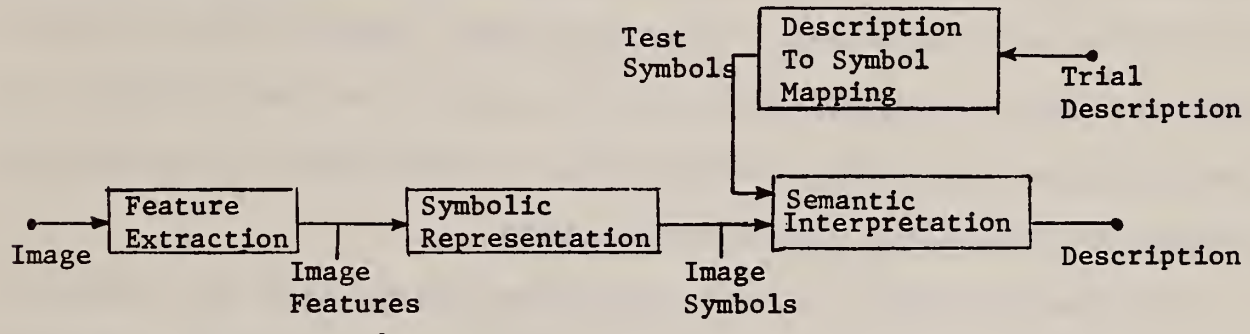
*This chapter is primarily based on Pratt, 1978, pp. 570-574.

Figure 5

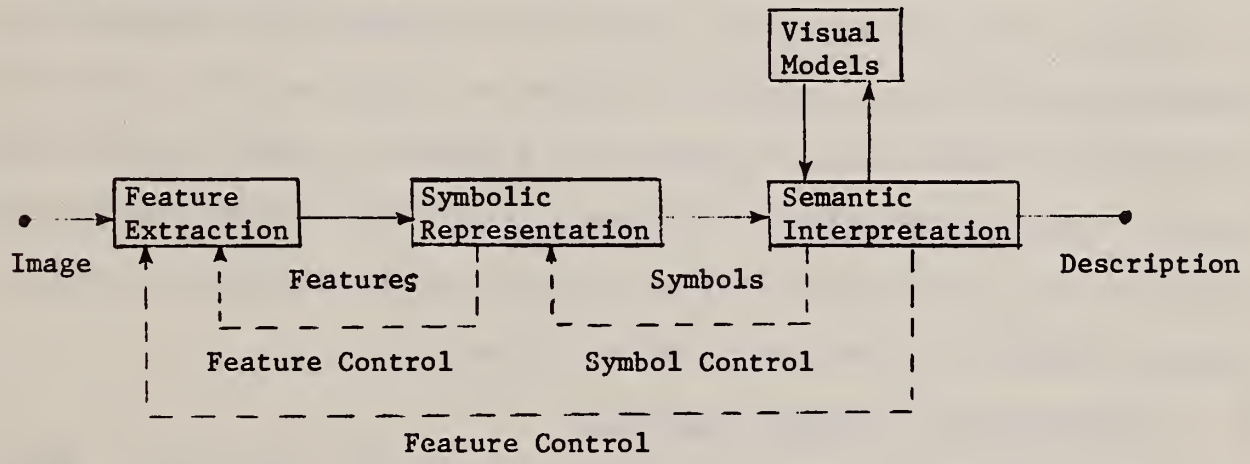
Basic Image Understanding Paradigms



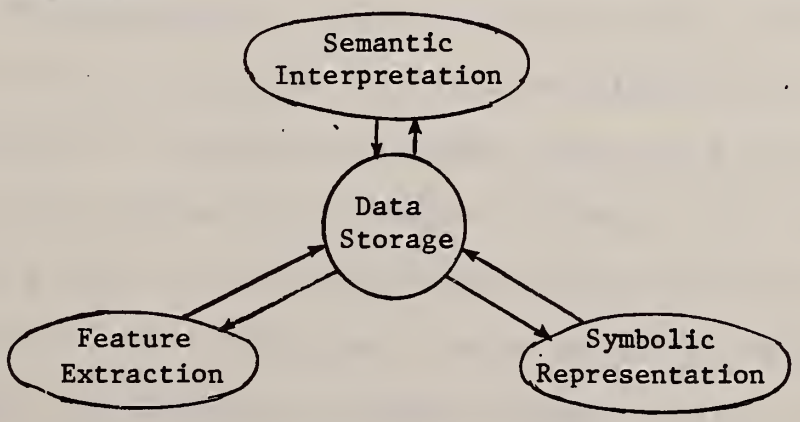
A. Hierarchical Bottom-up Approach



B. Hierarchical Top-down Approach



C. Heterarchical Approach



D. Blackboard Approach

Source: Pratt, 1978, pp. 570-574.

image. Examples of such features are picture element ("pixel") amplitudes, edge point locations and textural descriptors.

Next this set of features is passed on to the semantic interpretation stage where the features are grouped into symbolic representations. For example, edge points are grouped into line segments or closed curves, and adjacent region segments of common attributes are combined. The resultant symbol set of lines, regions, etc., in combination with a priori stored models, are then operated upon (i.e., semantically interpreted) to produce an application dependent scene description.

Bottom-up refers to the sequential processing and control operation of the system starting with the input image. The key to success in this approach lies in a sequential reduction in dimensionality from stage to stage -- vital as the relative processing complexity is generally greater at each succeeding stage. The hierarchical bottom-up approach can be developed successfully for domains with simple scenes made up of only a limited number of previously known objects.

B. Hierarchical Top-down Approach

This approach (usually called hypothesize and test), shown in Figure 5B, is goal directed, the interpretation stage being guided in its analysis by trial or test descriptions of a scene. An example would be using template matching -- matched filtering -- to search for a specific object or structure within the scene. Matched filtering is normally performed at the pixel level by cross correlation of an object template with an observed image field. It is often computationally advantageous, because of the reduced dimensionality, to perform the interpretation at a higher

level in the chain by correlating image features or symbols rather than pixels.

C. Heterarchical Approach

Hierarchical image understanding systems are normally designed for specific applications. They thus tend to lack adaptability. A large amount of processing is also usually required. Pratt (1978, pp. 572-573) observes that often much of this processing is wasted in the generation of features and symbols not required for the analysis of a particular scene. A technique to avoid this problem is to establish a central monitor to observe the overall performance of the image understanding system and then issue commands to the various system elements to modify their operation to maximize system performance and efficiency.

Figure 5C is a block diagram of an image understanding system that achieves heterarchical operation by distributed feedback control. If the semantic interpretation stage in the model experiences difficulty in working with its input symbol set, control can be fed back to the symbolic representation stage to request a new set of symbols. This action in turn may result in a command to the feature extraction stage requesting a modified set of features. When required, direct feedback control is also possible between the semantic interpreter and feature extractor. This paradigm provides an important auxiliary benefit in addition to flexibility. That is, the dimensionality of the feature and symbol sets can be kept at minimum levels because the sets can be restructured on command.

D. Blackboard Approach

Another image understanding system configuration called the blackboard model has been proposed by Reddy and Newell (1975).

Figure 5D is a simplified representation of this approach in which the various system elements communicate with each other via a common working data storage called the blackboard. Whenever any element performs a task its output is put into the common data storage, which is independently accessible by all other elements. The individual elements can be directed by a central control, or they can be designed to act autonomously to further the common system goal as required. The blackboard system is particularly attractive in cases where several hypotheses must be considered simultaneously and their components need to be kept track of at various levels of representation.

VIII. Levels of Representation

A computer vision system, like human vision, is commonly considered to be naturally structured as a succession of levels of representation. Tenenbaum et al. (1979, pp. 242-243) suggests the following levels (listed from low to high):

Images
Pictorial features
Intrinsic surfaces and bodies
3-D surfaces and bodies
Space map
Symbolic relationships

Tenenbaum et al. contrast this with current industrial vision systems relying heavily on detailed models of particular objects to accomplish tasks, employing levels of:

Images
Pictorial features (Edges & Regions)
2-D feature attributes
Objects (specific 2-D views)

Current industrial systems usually begin by thresholding the original gray-level image to obtain a binary array. Pictorial features (regions or edges) are then extracted from the gray-level or binary image and equated with surfaces or surface boundaries. These 2-D attributes of these pseudo-surface features are then symbolically matched against 2-D models (representing specific views of expected objects) to achieve recognition. As these industrial systems rely on prototype 2-D representations of anticipated objects, they are very limited for use in more general environments.

Barrow and Tenenbaum (1981, pp. 580-581) suggest the levels given in Figure 6 as those appropriate to a general-purpose vision system. The processing steps in the figure that transform each level of representation to the next require knowledge from

models of the physics of the imaging process, the illumination and the scene. At the lower levels, these models help resolve the ambiguity associated with going from a three dimensional world to a two dimensional image. At the higher levels, these models provide a foundation for organizing surface fragments into recognizable objects.

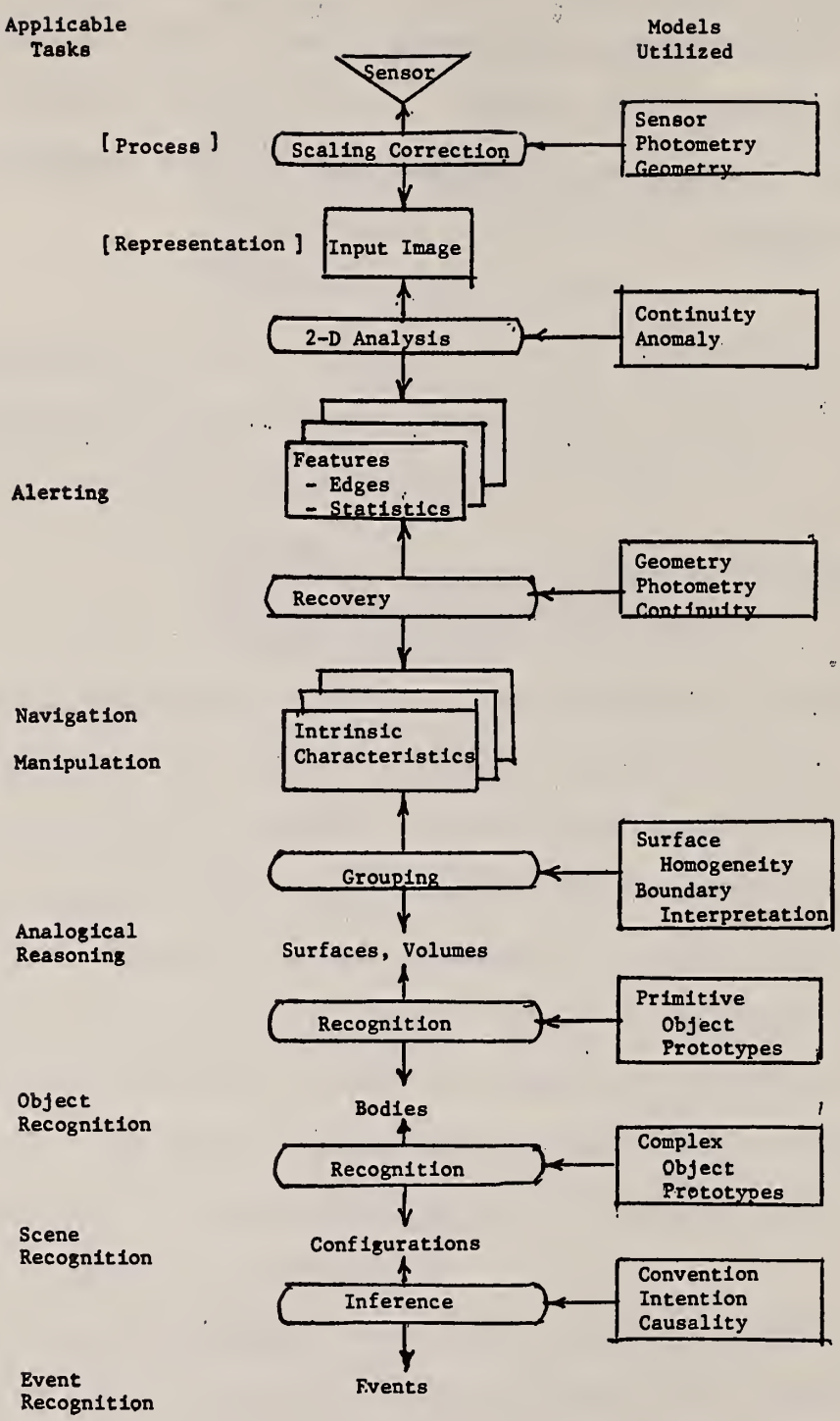


Figure 6: Computational Architecture for a General-Purpose Vision System

Source: Barrow and Tenenbaum, 1981, p. 58.

The input models required to do the processing at each level are shown at the right. On the left are shown the tasks for which vision can be used at each level of processing

Tenenbaum, et al., (1979, pp. 254-255), sketch in Figure 7 another way in which to view an organization of a vision system. They divide the figure into two parts. The first is image oriented (iconic), domain independent, and based on the image data (data driven). The second part of the figure is symbolic, dependent on the domain and the particular goal of the vision process.

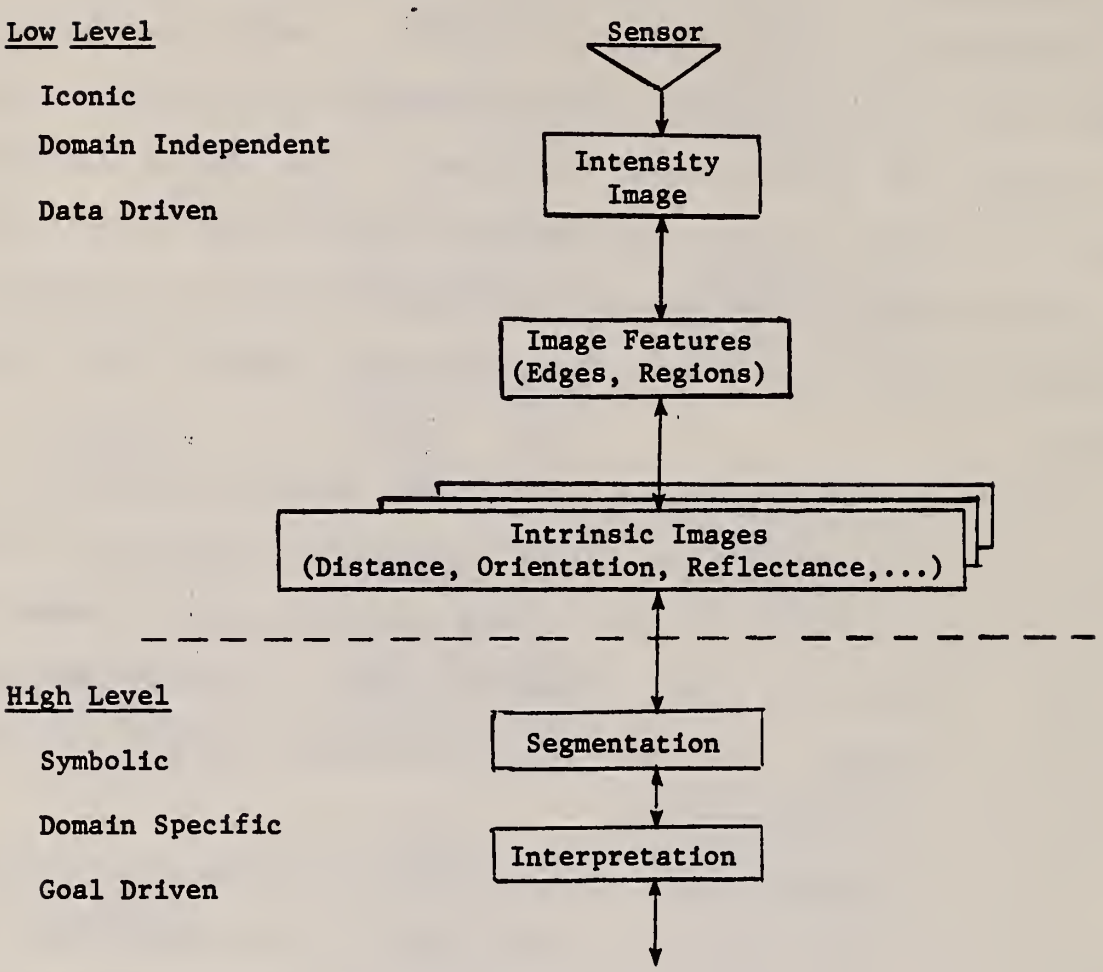
The first portion takes the image, which consists of an array of intensity of picture elements ("pixels," e.g., 1000x1000), and converts it into image features such as edges and regions. These are then converted into a set of parallel "intrinsic images", one each for distance (range), surface orientation, reflectance*, etc.

The second part of the system segments these into volumes and surfaces dependent on our knowledge of the domain and the goal of the computation. Again using domain knowledge and the constraints associated with the relations among objects in this domain, objects are identified and the scene analyzed consistent with the system goal.

*Fraction of normal incident illumination reflected.

Figure 7

Organization of a Visual System



Source: Tenenbaum et al., 1979, p. 255.

Reviews of the representation methods and techniques for performing the operations indicated in Figure 7 are given in the appendices.

The next chapter (Chapter IX) provides an overview of research in model-based vision systems. These systems endeavor to start with an image and produce, using a priori models, a desired description of the original scene, thereby spanning the complete hierarchy of Figure 7. The systems are constructed using the various representations, techniques and models reviewed in the appendices.

IX. Research in Model-based Vision Systems

Most research efforts in vision have been directed at exploring various aspects of vision, or toward generating particular processing modules for a step in the vision process rather than in devising general purpose vision systems. However, there are currently two major U.S. efforts in general purpose vision systems: The ACRONYM system at Stanford University under the leadership of T. Binford, and the VISIONS system at the University of Massachusetts at Amherst under A. Hanson and E. Riseman.

The ACRONYM system, outlined in Table II-1, is designed to be a general purpose, model-based system that does its major reasoning at the level of volumes rather than images. The system basically takes a hierarchical top-down approach as in Figure 5c. ACRONYM has four essential parts: modeling, prediction, description and interpretation. The user provides ACRONYM with models of objects (modeled in terms of volume primitives called generalized cones) and their spatial relationships; as well as generic models and their subclass relationships. These are both stored in graph form. The program automatically predicts which image features to expect. Description is a bottom up-process that generates a model-independent description of the image. Interpretation relates this description to the prediction to produce a three-dimensional understanding of the scene.

The VISIONS system outline in Table II-2, can be considered to be a working tool to test various image understanding modules and approaches. Rather than using specific models, its high level knowledge is in the form of framelike "schemas" which

represent expectations and expected relationships in particular scene situations. VISIONS is based on monocular images and does its reasoning at the level of images rather than volumes.

Other research efforts in model-based vision systems are summarized in Tables III in Appendix I.

It will be observed that each system is individually crafted by the developer to reflect the developer's background, interests and domain requirements. All, except ACRONYM (and to an extent MOSAIC), use image (2-D) models and are viewpoint dependent. Models are mostly described by semantic networks, though feature vectors are also utilized. The systems capitalize on their choice to limit their observations to only a few objects, by using predominantly a top-down interpretation of images, relying heavily on prediction.

Table II-1

Model-Based Vision Systems

Developer: Brooks et al. (1979), Brooks (1981)
System: ACRONYM
Purpose: General Purpose Vision System
Example Domains: Identifying Airplanes on a Runway in Aerial Images
Simulation for Robot Systems and for Automated Grasping of Objects

Approach	Modeling	Image Feature Extraction & Representation	Search & Matching	Remarks
Hiearchical top down approach. Reasons between different levels of representation based on a hierarchy of representations. High level modeler provides a high level language to manipulate models using symbolic names. Predictor and Planner Module is a rule-based system to generate an Observability Graph from the Object Graph (3-D object representation consisting of nodes and relational arcs). Makes predictions (which are viewpoint insensitive) in the form of symbolic constraint expressions with variables. Makes a projective transformation from models. Predicts appearances of models in images in terms of ribbons and ellipses. Incorporates translation and rotation into observable representations. Searches for instances of models in images. It employs geometric reasoning in the form of a rule based problem-solving system. It interprets (matches) in 3-D by enforcing constraints of the 3-D model.	Represents object classes from which subclasses and specific objects are represented by numeric constraints Models 3-D objects using volume primitives: generalized cones and ribbons. Spatial relations of volume elements within an object defined hierarchically. Can model both specific and generic volume elements and relations between them. Models are part/whole graphs Volume primitives have local rather than viewer-centered primitives.	Ribbons and curves obtained from an edge mapper. Surfaces obtained from a stereo mapper. Nodes of the Picture Graph (symbolic version of image) correspond to ribbons, surfaces and curves. Arcs and relations indicate spatial relations between nodes.	Matcher does an interpretation matching by mapping the Observability Graph into the Picture Graph. Matcher works in a coarse to fine order. Combines local matches of ribbons into clusters. Searches for maximal subgraph matches in the Observability Graph. Performs major interpretation at the level of volumes rather than at the level of images.	Aims to be a general vision system. Insensitive to viewpoint. A goal is to make use of total information for interpretation. Feature extraction (e.g., finding lines and regions) still weak. Interpretation is limited to scenes with few objects. Substantial progress has been achieved in past few years.

Developer: Hanson & Riseman (1978b,c)

Systems: VISIONS

Purpose: Interpreting static monocular scenes
Can be considered to be a working tool to test various image understanding modules and approaches

Example Domains: House scenes from ground level
Road scenes from ground level

Table II-2
Model-Based Vision Systems

Approach	Modeling	Image Feature Extraction & Representation	Search & Matching	Remarks
Uses a hierarchical modular approach to representation and control. Tries to be as general as possible to allow both bottom-up and top-down solution hypotheses as well as various intermediate combinations Incorporates the flexibility to utilize various feature extraction modules and multiple knowledge sources as required Allows for the possibility of generating and verifying hypotheses along many paths	Hierarchical structure Scene schemas (like frames) are the highest representation Hierarchy is: -schemas -objects -volumes -surfaces Proposed representations of 3D surfaces and volumes include: -generalized cylinders -surface patches with cubic B-splines to represent boundary and blending functions Employs semantic networks -nodes represent primitive entities (objects, concepts situations, etc.) -Labeled arcs represent relationships between them	Uses both edge finding and region growing to segment the image into a layered directed graph of regions, line segments and vertices Uses a hierarchical processing cone (pyramid) to be able to handle image data at various levels of resolution Uses a relaxation approach to organize edges into boundaries, and pixel clusters into regions, using high-level system guidance (interpretation guided segmentation)	Generates and stores partial models in "contexts" (of the CONNIVER programming language) which provide a history of decisions to be used when backtracking is necessary Uses a multiple knowledge source heterarchical approach which generates partial models in the search space of models. Attempts, using top-down and bottom-up relaxation techniques, to converge on a most probable solution. Uses rules for focussing on an element of a task, expanding that element by generating new hypotheses and verifying new hypotheses.	System (Parma, 1980) did reasonably well in making a crude segmentation of a house scene Viewpoint dependent Schema used depends on specific scene.

X. Industrial Vision Systems

A. General Characteristics

The prominent aspect of industrial vision systems, in distinction to more general vision systems, is that they operate in a relatively known and structured environment. In addition, the situation (such as placement of cameras and lighting) can be configured to simplify the computer vision problem. Usually, the number and nature of possible objects will tend to be restricted, and the visual system will be tailored to the function performed. Thus many of them are based on a pattern recognition, rather than an image understanding approach. Industrial vision systems are characteristically used for such activities as inspection, manipulation and assembly.

In an inspection task, the focus is on deviations from a standard, and usually little or no information is needed for identification. A manipulator controller, designed to pick parts off a conveyor, needs to be able to determine the identity, orientation and position of parts, but needs to know little of their precise shape except, perhaps at the grasp point. A visual controller for an arc welder will have its focus on the seam properties and needs little information about the appearance of the parts.

Kruger and Thompson (1981, p. 1525), in discussing the design of industrial vision systems, state:

> The complexity of most perceptual tasks requires that the problem be decomposed into manageable subunits. Thus major design decisions include the function of each module, the computational techniques and data representations imbedded in each module, and the control structures that relate modules and transfer

information between them. Most computer vision systems use a hierarchical organization...

A popular organization for industrial computer vision is a two-stage hierarchy with a bottom-up control flow. The lower level segments the image into regions corresponding to object surfaces. The higher level uses this segmentation to identify objects from their surface descriptions.

In practice, most successful systems incorporate aspects of both bottom-up and top-down control. The bottom-up processing is used to extract prominent features of a part to determine its position. Then, top-down control is used to direct a search to determine if the part satisfies an inspection criterion.

Industrial inspection and assembly operations are well suited to model-based analysis, because of the well-defined geometric descriptions associated with manufactured items. CAD/CAM technology allows the specification of objects using either volumetric or surface-based models. These geometrically based models are particularly appropriate to the hypothesis-verify approach, in which low-level image features are extracted and matched to an appropriate computer generated 2-D representation.

In addition to geometric models, objects may also be represented by graphs. In this case, recognition becomes a graph-matching process.

More commonly at present, rather than using geometric models or graphs, industrial vision systems are taught by being presented sample parts to be recognized in each of their expected stable states. Aspects of the resulting images are typically

stored as templates, and recognition becomes template matching. The objects can also be represented in terms of their characteristic features, such as area, number of holes, etc., and the resulting feature vector stored to be matched (via a search process) to the corresponding extracted feature vector of the image during system operation.

To simplify industrial vision systems, the input is usually reduced to a binary (black and white) image, so that objects appear as silhouettes. Simplicity is important in industrial vision systems because the computation time is limited, as most systems are expected to operate in near real time.

B. Examples of Efforts in Industrial Visual Inspection Systems

Table IV (based largely on Kruger and Thompson, 1981) lists some example efforts of vision systems designed for inspection. The systems listed are primarily for the inspection of printed circuit boards and IC chips, with template matching being the predominant inspection approach.

Kruger and Thompson (1981, p. 1529) note that: "Automated visual analysis has also been applied to the inspection of surface properties such as roughness, scratching and other potential defects. The best successes have come with highly specialized illumination and sensing systems, specifically tailored for a particular application. Recently, greater sophistication in the modeling of the imaging process has lead to prototype surface inspection systems with the promise of increased generality."

Chin (1982) has recently published an extensive bibliography on automated visual inspection techniques and applications.

Table IV

Example Efforts In Industrial Visual Inspection Systems

Developer Purpose Sample Domains	Approach	Modeling and Representation
Baird (1978) Inspection Automated manufacture of power transistor-pair IC chips GM	Inspection process consists of 1) Detection of the IC location and orientation on the heat-sink substrate 2) Quality control assessment after acquisition A gradient edge detector is used to compile the histogram of all edge directions in the inspection field. Peak of this histogram indicates the approximate orientation of the chip Next the corners of the IC are located by template matching. If any corners not located, the IC rejected Cracked, fractured chips are eliminated by a simple contrast thresholding operation	Inspection field consists of a 50x50 pixel region digitized to 16 gray levels Templates for IC corners
Chin, Harlow & Dwyer (1977) Inspection PCB's U. of Maryland	<u>Training Phase</u> Use an operator-interactive model-building graph procedure to train the inspection system Using an interactive camera/display system, the binary image edges of of a prototype PCB board are detected, smoothed to reduce noise, and encoded into a compact data structure. <u>Inspection Phase</u> Matching (against prestored edges and the graph model) is used to detect flaws in test images	Edges, Graph Model
Krakauer & Pavlidis (1979) Inspection Mass-Produced PCB's Princeton University	Ingenious use of binary template matching using a limited number of well-chosen templates accessed via a rapid lookup technique	Binary Templates
Jarvis (1980) Inspection Mass-Produced PCB's Bell Labs	1) Local pattern matching to stored binary templates 2) Supplemental tests for suspicious regions -Computation of conductor area -Length of the conductor-substrate boundary -ratio of area to length Processing done with simple logical operations	List of 5x5 pixel binary templates
Hseih & Fu (1979, 1980) Inspection and wirebonding guidance Multi-layered IC chips Purdue University	Inspection paradigm for proposed system is (for the most part) top-down and model-driven using a tree-like syntatic approach Various inspection algorithms are called for based on the actions of a controller, which monitors the whole vision process First, the image goes thru a series of task and context-dependent filters to reduce ambiguities Then, 8 special purpose defect detectors are used, as required	Design and inspection specification take the form of a descriptive data base Six subpattern masks

C. Examples of Efforts in Industrial Visual Recognition and Location Systems

Table V (again largely derived from Kruger and Thompson, 1981) lists some example efforts of vision systems designed for industrial part recognition and location. All these systems use a bottom-up approach. It will be observed that (except for Vamos 1979, and Albus et al., 1982) these systems utilize template or feature vector matching. Vamos does work from a 3D wire frame model which utilizes computer graphics type techniques to transform a model projection into alignment with observed lines in the image.

Albus' Machine Vision Group in the NBS Industrial Systems Division is using simplified 3D surface models of machined parts to generate expectancy images from needed viewpoints. The group is seeking to achieve real-time, hierarchical, multi-sensory, interactive robot guidance.

D. Commercially Available Industrial Vision Systems*

Table VI in Appendix J lists many of the Industrial Vision Systems that are currently commercially available. Most of the systems require special lighting.

It will be observed that many of the systems designed for verification and inspection use pattern recognition, rather than AI techniques. The systems tend to be bottom-up because of the speed requirements to achieve real time operation. Often unique edge and feature extraction algorithms are programmed in hardware or firmware.

*Additional information can be found in Gevarter (1982A).

Table V

Example Efforts in Industrial Visual Recognition and Location Systems

Developer Purpose Sample Domains	Approach	Modeling and Representation
Agin (1980), SRI SRI Vision Module Locate, identify and guide manipulation of industrial parts Engine Parts	Bottom-up approach Uses thresholding to convert to a binary image Each line is sequentially scanned and edge points (where pixels change form 1 to 0 or 0 to 1 recorded). Each resulting segment on a line is matched to the previous line to determine their overlapping relationships. Using these relationships, the program traces the appearance and disappearance of blobs (regions) as the image is processed from top to bottom. Using blob descriptors, the system can recognize parts regardless of their position or orientation. The descriptors are matched using either a binary decision tree or a normalized nearest-neighbor method. The system is trained by repeatedly showing the object to the TV camera resulting in all potentially useful shape descriptions being automatically calculated and stored	Blob descriptors include: -max. and min. x and y values -Holes -Area -Moments of inertia -Perimeter length -Linked list of coordinates on the perimeter
Kashioka et al (1976), Hitachi Central Research Lab Location and Bonding Guidance Transistor wire-bonding	Template matching Locates appropriate base and emitter leads on a semi-conductor chip so that wires can be stretched and bonded between them Initially trained by man-machine interactive selection in the universe of templates Multiplexed computer architecture to accommodate separate cameras on up to 50 bonding machines on a time-shared basis	Local 12x12 pixel binary templates

Table V (continued)

Example Efforts in Industrial Visual Recognition and Location Systems

Developer Purpose Sample Domains	Approach	Modeling and Representation
Holland Rossol & Ward (1979) Consight I Industrial part location, recognition and manipulation Engine parts GM	Two linear light sources superimpose a line of light on a conveyor belt perpendicular to its direction of motion. The two lines separate, proportional to the part passing by. Point of separation determines part boundary; degree of separation determines part thickness. The scene is imaged with a linear array camera and a silhouette automatically generated. Uses same feature vector approach as SRI Module.	Feature vector of part image characteristics
NBS: Albus et al. (1982) Visual servoing for robot guidance (real-time location and identification for manipulation) Machined parts National Bureau of Standards	Employs a point light source, a sheets-of-structured-light generator and a camera, all mounted on the wrist of a robot arm. Uses alternate frames of: 1. A regular point source illumination of the entire object, and 2. Two parallel planes of structured light. System determines location and orientation based on triangulation (associated with relative height of intersection of light sheets with part), and recognition based on shape and size of observed lines that the planes of light makes as it intersects part. Uses this information to interpret outline seen in image produced by the point source illumination. Analysis of vision input is performed with a hierarchically organized group of microprocessors. At each level of the hierarchy, and analytic process is guided by an expectancy-generating modeling process. The modeling process is in turn driven by a store of a priori knowledge, by knowledge of the robot's movements, and by feedback from the analytic process. Each such level of the hierarchy provides output to guide a corresponding level of the robot's hierarchical control system.	Uses quadratic approximations to surfaces of idealized 3-D objects.
Perkins (1978) Industrial parts recognition Engine components GM	Operates on 32 gray levels Bottom-up scene segmentation approach 1. Reduce 256x256 pixel image to an "edge gradient" image 2. Link edges with similar gradient magnitudes to form chains 3. Characterize chains as either straight lines or circular arcs. (This reduces 65,000 pixel image to about 50 concurves.) System matches observed concurves with model generated concurves using: 1. A preset control structure to select the order in which combinations of model and scene concurves are to be matched. 2. Starts by matching one model and one scene concurve 3. The stored model is spatially transformed and rotated to fit associated scene concurves System interactively trained by generating concurves of sample parts Can identify parts partially occluded by other parts	Concurve models of sample parts
Yachida and Tsuji (1978) Industrial Parts Recognition Nonoccluded parts of a small gasoline engine Osaka Univ.	Uses a boundary detection and isolation of parts in a binary image approach similar to SRI Vision Module Recognition system based on a structured step-by-step analysis with the previously stored models Use a series of special feature detectors -hole detector -line finder -texture detector -small hole detector System training involves interactive man-machine examination of the identification task	Stable orientation models of parts -part name -orientation -list of primitive features -polar coordinate boundary representation
Vamos (1979) Recognition of 3D Objects Bearing housings Assembly Sheet metal parts to be painted Neural nets in microscopic-section in neural research Hurgarian Acad. of Science	Finds edges using a simplified version of the Hueckel-operator using only two linear templates Lines are then fitted to edges Wire-frame model transformed (and hidden line elimination used) to correspond to image — yielding recognition and part orientation Objects are interactively taught to system either by building a geometric model or by a computer-aided transformation of viewed samples	3D Wire Frame Models

The more sophisticated systems tend to utilize variations and improvements on the SRI Vision Module described in Table V.

A few systems make good use of structured light for 3D sensing. A number of efforts in guidance of arc welding take this form.

XI. Who is Doing It

Rosenfeld, at the University of Maryland, issues a yearly bibliography, arranged by subject matter, related to the computer processing of pictorial information. The issue covering 1981, (Rosenfeld, 1982) includes nearly 1000 references.

The following is a list by category of the U.S. "principal players" in computer vision.

A. Research Oriented

1. Universities

These are shown in Table VII.

2. Non-Profits

SRI International, AI Center
JPL

3. U.S. Government

NBS, Industrial Systems Div., Gaithersburg, MD
NOSC (Naval Ocean Systems Center), San Diego.
NIH (National Institutes of Health)

B. Commercial Vision Systems Developers

A partial listing is given in table VIII. It has been reported that hundreds of companies are now involved in vision systems.

Table VII

University Organizations Engaged in Computer Vision Research

Artificial Intelligence and Computer Science Laboratories Funded Under DARPA IU Program

	C.S. Labs	A.I. Labs	Other
CMU			Robotics Institute
U of MD	X		
MIT		X	
U of Mass.			Comp. & Info. Sci. Dept.
Stanford U		X	
U of Rochester	X		
USC			Information Processing Institute
U of Rhode Island			Robotics Res. Lab
Other Active Universities			
U of Texas at Austin	X		
VPI	X		
Purdue	X		
U of PA	X		
U of IL	X		
Wayne State U	X		
JHU			E. E. Dept.
RPI			Elec. & Sys. Engr. Dept.

Table VIII

Commercial Vision System Developers

Industrial Vision Companies	Large Diversified Manufacturers*	Robot Manufacturers
Machine Intelligence Corp.	General Electric	Copperweld Robotics
Robot Vision Systems	Chrysler Corporation	Unimation
Videometrix	General Motors	Automatix, Inc.
Object Recognition Systems	Industrial Business Machines	
Octek, Inc.	Texas Instruments	
Cognex	International Harvester	
Spectron Engineering Inc.	Westinghouse	
Ham Industries	Hughes	
Quantomat	Lockheed - Palo Alto Research Lab.	
Image Recognition Systems	Fairchild Camera and Instrument Corp.	
Colorado Video	Martin Marietta	
Everett Charles	McDonald Douglas Automation Company	
Inspection Technology	Cheesebrough Ponds	
View Engineering		
Vanzetti		
Automated Vision Systems		
Perceptron, Inc.		
Vicom Systems, Inc.		
Cyberanimation, Inc.		
Reticon		

*Some Systems are for in-house use only

XII. Who Is Funding It

To date, the principal source of funding for computer vision research has been the U.S. Government, which is estimated to spend in the order of $10 million a year in this area.

The major U.S. Government program has been the DARPA Image Understanding Program. Other government agencies funding vision research are:

NSF (National Science Foundation)
NIH (National Institutes of Health)
NBS (National Bureau of Standards)
ONR (Office of Naval Research)
DMA (Defense Mapping Agency)
NASA (National Aeronautics and Space Administration)
USGS (U.S. Geological Survey)
AFOSR (Air Force Office of Scientific Research).

It is estimated that DARPA (Defence Advanced Research Projects Agency) spends in the order of $2.5 million dollars a year in computer vision research. DARPA thrusts include automatic stereo and terrain mapping, autonomous navigation, robot vision, symbolic representation, autonomous expert image systems, and photo analysis aids. DARPA helps support a number of Image Understanding laboratories at universities where I.U. work at all levels is performed.

DMA has entered into a very active program in image and scene analysis. Their goal is to achieve "fully automated" production for mapping, charting and geodesy by 1995, in which the primary role of human beings will be to validate the inputs and the output extracted information. They intend to commence by furnishing computer vision aids to the cartographer and achieve the desired high-level automation via an evolutionary route. Their current approach is to focus a portion of DARPA's image

understanding effort on producing an Image Understanding Testbed for integrating and evaluating current and emerging computer vision techniques and systems. An initial version of this Testbed has been constructed at SRI in Menlo Park, CA (Hanson and Fischler, 1982). The future emphasis of the Testbed will be on expert systems that facilitate the application of IU research results to cartographic problems.

NSF spends roughly $1.5 million a year on a variety of research topics in computer vision.

NIH spends a substantial sum in obtaining and evaluating images for a variety of medical research applications. This has included efforts in semi-automatic cancer screening, computer-assisted photometry, tomography, image formation and imaging equipment and various other medical application related areas. As the focus is application oriented, rather than computer vision oriented, it is difficult to pinpoint the portion that can be considered computer vision research. However, a rough guess might put the figure at one to two million dollars a year.

It has been estimated that NASA spending on image processing and evaluation approaches one hundred million dollars a year. To help support this effort, NASA funds somewhat less than one million dollars a year on research in computer vision*. NASA spends roughly half this sum to support research at JPL in vision systems to guide robot manipulation.

*Additional funds have been spent on image processing and analysis hardware such as the Massively Parallel Processor (MPP) at the NASA Goddard Space Flight Center.

The National Bureau of Standards has an ongoing in-house robotics vision research effort, which currently is in the order of one half million dollars a year.

Collectively, other government agencies probably spend another one to two million dollars per year for research in this area. It is estimated that perhaps an additional one to two million dollars a year is spent by government contractors using IRAD (Independent Research and Development) funds associated with their prime contracts.

XIII. Summary of the State-of-the-Art

A. Human Vision

Human vision is the only available example of a general purpose vision system. However, thus far not many AI researchers have taken an interest in the computations performed by natural visual systems, but this situation is changing.

The MIT vision group (among others) believes that, to a first approximation, the human visual system is subdivided into modules specializing in visual tasks. There is also evidence that people do global processing first and use it to constrain local processing.

Considerable information now exists about lower level visual processing in humans. However, as we progress up the human visual computing hierarchy, the exact nature of the appropriate representations becomes subject to dispute. Thus, overall human visual perception is a still very far from being understood.

B. Low and Intermediate Levels of Processing

Though methods for powerful high-level understanding visual analysis are still in the process of being determined, insights into low-level vision are emerging. Alan Mackworth, from the University of British Columbia observed at IJCAI-81 that there is an exciting convergence in the theory of low level vision from the major vision centers, such as MIT, CMU, SRI and Stanford. The basic physics of imaging, and the nature of constraints in vision and their use in computation is fairly well understood. Detailed programs for vision modules, such as "shape from shading" and "optical flow," have begun to appear. Also, the representational issues are now better understood.

However, even for well understood low-level operations such as edge detection, there has been no convergence among the many techniques proposed, and no method stands out as the best. In general, edge detectors are still unreliable, though Marr and Hilbert's approach, based on the zero crossing of the second derivative of the intensity gradient, appears promising. Brady (1981B, p. 3) states that operators designed to extract the "important" intensity changes in an image are still more an art than a science. Approaches to edge detection consist mostly of convolving images with local operators tuned to particular applications. These operators fare badly outside their limited domain or in the presence of noise.

Barrow and Tenenbaum (1981, p. 576) note that the direct approach to image segmentation is inherently unreliable. A number of research groups successfully circumvented this problem by integrating segmentation and interpretation. However, this approach is not suitable for a general purpose vision system as it is based on advance knowledge of the objects to be expected.

In industrial vision, the primary technique for achieving robust edge finding and segmentation is to use special lighting and convert to a silhouette binary image in which edges and regions are readily distinguishable.

At intermediate levels, edge classification and labelling have been very successfully used in the blocks world. Barrow and Tenenbaum (1981, p. 573) believe that the various techniques developed for dealing with the blocks world could be integrated into a complete, highly competent vision system for that

domain. Thus far, however, no such system has actually been built.

Binford (1982) in reviewing existing research in model-based vision systems observed that most systems first segment regions then describe their shape. None of the systems makes effective use of texture for segmentation and description. In general, shape description is primitive and interpretation systems have not yet made full use of even these limited capabilities.

As yet, the extraction of useful information from color is extremely rudimentary. The perceptual use of motion (optical flow) has been a focus of attention recently, but findings are preliminary.

For low level processing, many recent algorithms take the form of parallel computations involving local interactions. One popular approach having this character is "relaxation." These locally parallel architectures are well suited to rapid parallel processing using special purpose VLSI chips.

C. Industrial Vision Systems

Barrow and Tenenbaum (1981, p. 572) observe that:

> Significant progress has been made in recent years on practical applications of machine vision. Systems have been developed that achieve useful levels of performance on complex real imagery in tasks such as inspection of industrial parts, interpretation of aerial imagery, and analysis of chest X-rays. Virtually all such systems are special purpose, being heavily dependent on domain-specific constraints and techniques. For example, industrial vision systems usually require high contrast to obtain binary images and use overhead cameras to minimize variations in object appearance.

A much more pressimistic view is taken by Kruger and Thompson (1981, p. 1524) who state that:

> Despite substantial research efforts, the study of computer vision is still in its infancy... Significant reductions in complexity are possible if automated perception is limited to an industrial environment. Even here, however, we still lack a clear understanding of the fundamental problems that must be addressed if computer vision is to have a major impact on manufacturing.

Hiatt (1981, pp. 2, 3) observes that in industry, robot vision systems are limited to simple repetitive processes, and that the classic bin of jumbled parts problem still overwhelms industrial vision systems. However, Birk and Kelley at the University of Rhode Island have devised algorithms to successfully pick out parts from a bin on up to 90% of the computer-vision robot's machine cycles.

Krueger and Thompson (1981, p. 1537) observe that, "The current state of the art precludes the construction of one general-purpose computer vision system with applicability to all industrial vision tasks.... Current systems use no common primitives for formal representations of object properties. There is also no common programming language for these applications. [Current industrial vision systems are limited in their flexibility in allowing users to reprogram the system to new situations.] This situation will likely improve as computer vision becomes more integrated into the production process."

In adapting concepts generated in the research laboratory to industrial vision applications, many important additional factors come into play such as speed, cost and complexity. It has also been found that the lighting and optics play a key role in the robustness of an industrial system. Most potential industrial vision applications cannot be reduced to working with binary

silhouettes, due to texture and other real-life environmental factors. Thus, systems engineering is an important ingredient. Unfortunately, at present many prospective users have inadequate inhouse capability to do the systems planning and integration needed to successfully adapt computer vision to their operations. This has inhibited the industrial use of sophisticated vision systems. The vision manufacturers are now beginning to try to remedy this situation by starting to provide easier user programming, friendlier user interfaces, and systems engineering support to prospective users.

It has been estimated that as of mid-1982, though less than 50 sophisticated industrial vision systems were actually in use, approximately 1000 simple line-scan inspection systems were in regular operation. Though special purpose systems have thus far been the most effective, successful vision applications are now becoming commonplace and are expanding. Many firms are now entering the industrial vision field, with technical leap-frogging being common due to rapidly changing technology.

D. General Purpose Vision Systems

1. Introduction

Though many practical image recognition systems have been developed, Hiatt (1981, pp. 2, 8) observes that, "In current vision applications, the type of scene to be processed and acted upon is usually carefully defined and limited to the capability of the machine... General purpose computer vision has not yet been solved in practice." This domain specificity makes each new application expensive and time consuming to develop. Thus, there is a clear need for computer vision systems capable of dealing

with a variety of industrial applications, particularly those with less structured environments.

Barrow and Tenenbaum (1981, p. 572) note that "Developing general-purpose computer vision systems has proved surprisingly difficult and complex. This has been particularly frustrating for vision researchers, who daily experience the apparent ease and spontaneity of human perception. Research in the last few years, however, has provided new insights into the computational nature of vision that could lead to systems capable of high performance in a broad range of visual domains."

Brady (1981A) observes that there has been a research shift toward topics corresponding to identifiable modules in the human vision systems, and away from particular domains of application. The consequence has been a sharp decline in the construction of entire vision systems.

2. Difficulties

Barrow and Tenenbaum (1981, p. 574) emphasize that

> Model-based interpretation of image data is an enormously complex computational task. The variety of possible scene configurations and viewpoints is so great that an exhaustive search through the space of possible interpretations is out of the question. Only the most promising or most important alternative interpretations can be pursued. Selection of candidate interpretations depends both upon information derived from the input image, and upon the observer's goals and expectations. A delicate balance must be struck between data-directed and goal-directed search to avoid oversight (not seeing things that are really present) and hallucination (seeing things that are not).

Gennery et al. (1981, pp. 10-1, 10-3, 3-6) observe that

> The statement "Vision is hard" is found often in the computer vision literature. There are several reasons for the difficulty. In the first place, an image contains an enormous amount of information, much

> of it irrelevant to the task at hand, and it is an imperfect projection of the real world, containing noise and distortion. From this the relevant information must be extracted. In the second place, the transformation from the image to the real world is highly ambiguous. Thus world knowledge must be relied on to resolve the ambiguities. (This is especially true in monocular vision of three-dimensional scenes, but it is also true to a lesser extent in stereo vision.) In the third place, an object seen may only vaguely resemble others of its generic type or even itself at other times or under other conditions. In the fourth place, in a powerful vision system an object must be recognized out of a large number of possible objects or generic types.
>
> These facts appear to manifest themselves in two ways in practice. First, vision requires an enormous amount of computing. Second, it seems that the computational methods needed are very complicated, and it is unknown today what the right methods will be...
>
> Some experimental systems hold promise for recognition of generic three-dimensional objects, although they require a large amount of computing time on existing computers. Some special-purpose hardware is becoming available, which enables some very low-level computations to be performed rapidly. Even in these cases, however, a variety of techniques are in use, with no consensus about which are the best. This becomes even truer as we move to the higher-level, more general, on more advanced areas. Furthermore, many of the approaches that have been used are ad hoc, with little promise of generality.
>
> ...two tasks that are beyond the capability of any existing computer vision system are the recognition of parts in a jumble in a bin and the operation of a robot vehicle in a complicated outdoor environment.

Rosenfeld (1981, p. 3) observes that "Image processing and scene analysis have definitely saturated the capacity of computers."

In relation to earth observation imagery for resources management, Alan Mackworth of UBC stated at IJCAI-81 that it will be necessary to alter the popular current multi-spectral paradigm that pixel meaning can be determined by intensity alone -- it doesn't work. It is necessary to understand spatial

organization, meaning and context. The spatial constraints are very important. There is no chance of getting a general purpose vision system to understand satellite imagery alone -- it is necessary to use a system-generated "sketch map" to interact with the scene.

3. Techniques

Brady (1981B, p. 4) observes that, "Most AI workers have... abandoned the idea that visual perception can profitably be studied in the context of a priori commitment to a particular program or machine architecture." Binford (1982) believes "...that building a vision system is 1% a system effort of the sort which are familiar in computer science, and 99% basic science."

The research emphasis has moved to developing techniques (vision modules) for extracting intrinsic images (shape from shading, shape from texture, etc.). Brady (1981A, p. 6) observes that, "Representations have been developed that make explicit, the information computed by a module... [This] leads to a view of visual perception as the process of constructing instances of a sequence of representations."

Gennery et al (1981, p. 3-1) note that at higher levels of descriptions it becomes difficult to judge what are the best approaches. As a result, a wide variety of techniques have been used.

Brady (1981, p. 99) observes that though it appears that the most difficult visual problem is the perception or planning of

movements through cluttered space, a solid start has been made on this problem by Lozano-Perez (1981).

4. Conclusions

Binford (1982) in reviewing current model-based research vision systems concludes that most systems have not attempted to be general vision systems, though ACRONYM does demonstrate some progress toward this goal. Existing vision systems performances are strongly limited by the performance of their segmentation modules, their weak use of world knowledge and weak descriptions, making little use of shape. The systems primarily relate image relations to image observables; in general lacking the ability to relate three dimensional space models to images. Existing systems show little emphasis on basic vision problems in systems building.

Binford observes that until recently, systems efforts have been small and short-lived, generally only a few man years effort. Focussed and continuous efforts are necessary but not sufficient for system building. The system programming effort alone in building a vision system is enormous.

With the exception of ACRONYM (and to an extent MOSAIC), the systems surveyed depend on image models and relations, and therefore are strongly viewpoint-dependent. To generalize to viewpoint insensative interpretations, would require three-dimensional modeling and interpretation as in ACRONYM.

Binford found that the systems jump to conclusions based on flimsy evidence which would probably not distinguish many objects in a complex visual environment. The systems typically use the hypothesis-verification paradigm. Hypothesis generation is the

crucial part, made easy in the top-down case. The systems succeed best with quasi-2d scenes, for example aerial photographs, industrial scenes from a fixed viewpoint, x-ray images, and ground level photos from a fixed viewpoint. Even ACRONYM, which incorporates viewpoint-insensitive mechanisms, has been demonstrated only on aerial images, although it appears applicable to ground level photographs as well.

Binford concludes that though that the results of these and other efforts are encouraging as first demonstrations, nevertheless as general vision systems, they have a long way to go.

Tenenbaum and Barrow (1981, p. 594) in discussing the general computer vision problem conclude that:

> We are beginning to understand the computational nature of vision at a fundamental level, independent of implementation. This understanding provides new insights into limitations of early scene analysis systems and a solid scientific foundation upon which future general-purpose high-performance computer vision systems can be built...
>
> The competence of a vision system ultimately rests upon the representations it uses to describe the world and the models available for manipulating and transforming descriptions. Many levels of description are necessary to achieve human performance requiring models of scene domains, objects, surfaces, illumination, sensors, and the geometry and photometry of imaging.
>
> A vision system is naturally structured as a sequence of levels of representation. The initial levels are primarily iconic (edges, regions, gradients) because that is the nature of the information available directly from an image. The highest levels are primarily symbolic (surfaces, objects, scenes), because that is the nature of the information that is sought. Intermediate levels are constrained by the information available from preceding levels and that required by subsequent levels. In particular, physical and three-

dimensional surface characteristics provide a critical transition from iconic to symbolic representations.

> Early levels of processing in a vision system are primarily data driven, while higher levels are controlled by goals and expectations. At intermediate levels, some combination of data-driven (bottom-up) and goal-driven (top-down) operation is needed both to compensate for errors, and to avoid computational overload. Although the detailed nature of processing is dependent on representation and therefore considerably different at low and high levels, it is significant that at virtually all levels processing appears to be inherently parallel, and thus amenable to implementation by networks of computational elements (e.g., neurons or VLSI chips)...
>
> While no such [general-purpose computer vision] system yet exists, most of the pieces have been experimentally demonstrated. Thus it would not be unreasonable to attempt to construct one within the current state of the art. Of course, many details still remain unresolved, especially at the higher levels of processing.

E. Visual Tracking

Real-time tracking of objects is important to manipulation and guidance. The state-of-the-art in visual tracking is reviewed and Appendix H. Though some success has been achieved under limited conditions, it remains as an important area for research.

F. Overview

In conclusion, we might observe that computer vision can be viewed as a set of very difficult problems. However commercial vision systems are available and are operating successfully in specialized environments on low level problems of verification,

inspection, measurement, recognition, and determination of object location or orientation.

There is now a much better understanding of the computer vision problem than there was just a few years ago. A major focus of the current research effort is in extracting 3D shape from intrinsic image characteristics.

Though quite a number of high level research vision systems have been explored, no general vision system is available today or is imminent. Major current efforts in this area are ACRONYM at Stanford U. and VISIONS at the U. of Mass.

XIV. Current Problems and Issues

A. General

Some of the general issues are:

- Can general vision be reduced to computer analysis?
  - What assumptions about the world are restrictive enough?
  - How much data is required?
- Need to incorporate generic aspects of perception (Binford, 1982)
  - Similarity, not spatial congruence is the paradigm of interpretation in nature.
  - Humans are always seeing things they haven't seen before.
- 3D interpretation of images versus 2D
  - Binford (1982) believes that general vision systems depend on building three dimensional descriptions--that prediction, description and interpretation take place largely in three dimensions.
- How necessary is it to follow the central paradigm (Figure 6, in this report) to achieve high level vision? Is it essential to employ a key intermediate representation such as the 2-1/2D sketch or Intrinsic Images? It is possible to obtain these using only local constraints?
- Is the hierarchical vision paradigm (Figure 6), which implies complete segmentation and labeling inappropriate for natural scenes? Is a more isomorphic representation needed, such as a map which implicitly captures the detail and relations and is more appropriate for natural computations? For such isomorphic representations, is the serial digital computer inappropriate and another calculating medium such as a network needed? (See Fischler, 1978, 1981.)
- Methods and hardware to reduce the software generation costs and processing time for computer vision*.

*Nudd (1980) provides a good overview of computer processing requirements for computer vision and appropriate architectures and hardware to implement them.

- Lack of interface standards for connecting computer vision systems to robots and industrial machines.
- Active vs passive sensing in vision systems.
- Relative merits of binary versus gray-scale imagery.
- Most issues are still poorly understood.

B. Techniques

1. Low Level Processing

- Many of the unsolved problems in computer vision are at this level.
- The whole issue of constructing the primal sketch from zero crossings of the second derivative of the intensity is far from resolved.
- Direct edge finding and region segmentation are still unreliable for general vision.
- A key insight is that local information is usually inadequate to guide segmentation and interpretation in a general scene. Global structure such as the shading gradient is required. To what extent can modeling and using physical contraints in scene analysis provide global restrictions which can guide segmentation and assist in classification? Possible examples include 1) utilization of shadows to locate lighting sources and to pinpoint objects casting shadows, and 2) use of sky-land boundaries as a global constraint. Another global approach for man-made scenes is to employ the camera model and geometric perspective to detect vanishing points associated with parallel lines in an urban scene. Fishler et al. (1982) indicate that the detection of clusters of parallel lines by finding their vanishing points can be used to automatically screen large amounts of man-made structures.
- How to best utilize and avoid difficulties with texture in natural scenes is still unsolved.
- Rectification of images prior to stereo matching remains a problem.

2. Middle Level Processing

- A key problem remaining in computer vision is bridging the gap between pictorial features (e.g., edges and regions) and 3D objects.

- Techniques for analyzing time-varying imagery.

- Limits of the intrinsic image approach: It is not clear that we can reliably obtain intrinsic images from images of real scenes via the methods outlined in this report. Alternative approaches when available, such as stereo or active ranging sensors, may be preferable for extracting intrinsic characteristics.

- How best to deal with shadows and occlusions.

3. Higher Level Processing

- Relation of higher level vision to AI

- Modules that operate on the surface orientation map to produce object representations.

- Generic interpretation in terms of object classes.

- Semantic interpretation.

- Semantic search techniques for use in matching schemes using semantic segmentations and indexing.

- Identification for interpretation of which geometric parameters are casually (functionally) rather than statistically determined (Binford, 1982).

C. Representation and Modeling

- Representations for complex and amorphous shapes (e.g., a tree, a crumpled sweater, a flowing stream).

- Proper level for the dividing line between iconic representations at the lower levels and symbolic at the higher levels, and how much these representations should overlap.

- How to index efficiently into a database containing a large number of models.

- What sort of features should be extracted from the scene (edges, corners, regions, surface orientation, etc.) and how should objects be modeled (wire-frame models, generalized cylinders, etc.)

D. System Paradigms and Design

- Is the relaxation process the most attractive approach at the lower levels where global aspects are not directly considered?

- How far can parallel methods (like relaxation) be pushed at all levels?
- Is a combination of top-down and bottom-up the preferred approach for complex vision tasks?
- Under what circumstances is the blackboard approach to be preferred? Note that hierarchical image understanding systems suffer from lack of adaptability and also require a large amount of processing.

E. Knowledge Acquisition - Teaching and Programming

- Methods for knowledge acquisition at all levels.
- Methods for learning and tracking of generic types.
- How to make the system versatile by having it programmable at a very high level.
- Design of a very-high-level programming language especially for vision. Little has been done to date in this area.

F. Sensing

- Active vs passive sensing.
- Best methods for using structured light.
- Methods for acquiring 3D directly.
- Is scanning laser radar the wave of the future?

G. Planning

- Methods to incorporate planning into robotic systems utilizing vision.

XV. Research Needed

A. General

- Need to understand human vision as it's our best example of a general purpose vision system.
- Need research in general purpose systems capable of high performance in a wide variety of visual domains.
- Need to be able to use generic recognition.
- Methods to reduce software costs of computer vision and reduce processing time.
- Image processing techniques for greater capability.
- Interface standards.
- Methods for visual guidance in cluttered spaces.
- Improved understanding of the extent and use of domain specific information in visual perception.
- Need to determine how to best utilize range information.
- Need to develop global methods (e.g., utilizing shading) either to bypass or to help guide the current hierarchical paradigm.

B. Techniques

1. Low level processing

- More reliable and faster edge and region finders for general scenes.
- Ways to extract motion measures from sequences of intensity arrays.
- Reliable stereo disparity modules.
- Determination of surface properties, such as color, smoothness, coatings, etc.

2. Middle level processing

- Techniques for analyzing time-varying imagery.
- Methods to bridge the gap between edges and regions and 3D objects.
- Improved methods of extracting intrinsic images.

3. High level processing

- Better understanding of how to use texture.
- Modules that operate on the surface orientation map to produce object representations.
- Methods for generic and semantic interpretation.

C. Representation and Modeling

- Representations for complex and amorphous shapes.
- Techniques for indexing into a large data base of models.
- Mathematical methods to model texture conviently.
- More precise representations of surface orientation maps at different levels of resolution.
- Methods to group properties at each level of resolution for each representation, so that a hierarchical structure can be imposed upon the representations.
- Determination of under what conditions is binary imagery most favorable and under what conditions is gray-scale to be preferred.
- Choosing which features to extract from a scene.
- Methods for modeling 3D objects.

D. System Paradigms and Design

- Need to explore the extension of relaxation processes to multiple levels in the pyramid of description and interpretation.
- Methods of local parallel processing which can discover global information through propogation.
- Efficient methods and techniques for maintaining concurrently a number of images of a scene in various stages of processing, so that these explicitly represented images can interact with each other and with higher and lower levels of processing as processing proceeds. This is especially pertinent to globally consistent relaxation processing.
- Need investigation of paradigms, other than the conventional hierarchical paradigms, such as the "blackboard" and other paradigms being explored in Artificial Intelligence areas such as expert systems.

- Faster pre- and post-processing hardware (e.g., special digital circuits to evaluate intensity gradients).

E. Knowledge Acquisition - Teaching and Programming

- Need better techniques for rapid reprogramming of vision systems by the user.
- Inspection and Assembly vision systems software approaches that can easily be modified to adapt to new situations. It would be desirable that control structures be incorporated that will specify tests to be performed and possible alternate paths of actions.
- Need high-level programming languages designed especially for vision.
- Need methods for learning and teaching of generic types.
- Methods for knowledge acquisition at all levels.

F. Sensing

- Techniques for rapid 3D sensing -- ranging by lidar (scanning laser radar), defocussing, stereo and triangulation.
- Improved methods for use of structured light for 3D evaluation and shape identification.
- Methods for exploiting multiple light sources.
- Higher-resolution and selective-resolution transducers.

G. Planning

- Methods for incorporating vision into robotic planning.

XVI. Future Trends

As the field of computer vision unfolds, we expect to see the following future trends*.

A. Techniques

- Though most industrial vision systems have used binary representations, we can expect increased use of gray scales because of their potential for handling scenes with cluttered backgrounds and uncontrolled lighting.

- Recent theoretical work on monocular shape interpretation from images (shape from shading, texture, etc.) make it appear promising that general mechanisms for generating spatial observations from images will be available within the next 2 to 5 years to support general vision systems.

- Successful techniques (such as stereo and motion parallax) for deriving shape and/or motion from multiple images should also be available within 2 to 5 years.

- The mathematics of Image Understanding will continue to become more sophisticated.

- Enlargement will continue of the links now growing between Image Understanding and Theories of Human Vision.

- Brady (1981B. p. 11) predicts that there will be a considerable advances in current vision "...issues over the next few decades, probably resulting in changes in our conception of computing and vision at least as large as those which have occurred over the past decade."

B. Hardware and Architecture

- We are now seeing hardware and software emerging that enables real-time operation in simple situations. Within the next 3 to 5 years we should see hardware and software that will enable similar real-time operation for robotics and other activities requiring recognition, and position and orientation information.

---

*These trends have been largely derived from statements by Brady (1981A, 1981B), Binford (1982), Kruger and Thompson (1981), Agin (1980), Arden (1980), Rosenfeld (1981), Hiatt (1981), and Barrow and Tenenbaum (1981).

- Fast raster-based pipeline preprocessing hardware to compute low-level features in local regions of an entire scene are now becoming available and should find general use in commercial vision systems in 2 to 4 years.

- As at virtually all visual levels, processing seems inherently parallel, parallel processing is a wave of the future (but not the entire answer). Parallel processing research hardware systems (such as ZMOB at the U. of MD, and the MPP for NASA Goddard) have already been built, and appropriate algorithms are being developed.

- Three possible parallel processing architectures are array processing, pipeline processing and multi-processing. Multi-processing looks most promising as it allows data from several data streams of an image to interact with each other to yield a high-level representation.

- Relaxation and constraint analysis techniques are on the increase and will be increasingly reflected in future architectures.

C. A.I. and General Vision Systems

Computer vision will be a key factor in achieving many artificial intelligence applications. The goal is to move from special-purpose visual processing to general-purpose computer vision. Work to date in model-based systems has made a tentative beginning. But the long-run goal is to be able to deal with unfamiliar or unexpected input*. Reasoning in terms of generic models and reasoning by analogy are two approaches being pursued. However, it is anticipated that it will be a decade or more before substantial progress will be made.

*As computer vision systems move toward this goal, they will increasingly incorporate Expert System components using multiple knowledge sources. Gevarter (1982B) provides An Overview of Expert Systems, in which ACRONYM and VISIONS are considered to be examples of Expert Systems.

Barrow and Tenenbaum (1981, p. 594) indicate that no general vision system now exists, but most of the "...pieces have been experimentally demonstrated. Thus it would not be unreasonable to attempt to construct one within the current state-of-the-art."

D. Modeling and Programming

- Now emerging is 3D modeling, arising largely from CAD/CAM technology. 3D CAD/CAM data bases will be integrated with industrial vision systems to realistically generate synthesized images for matching with visual inputs.
- Illumination models, shading and surface property models will be increasingly incorporated into visual systems.
- Volumetric models which allow prediction and interpretation at the levels of volumes, rather than images, will see greater utilization.
- High level vision programming languages (such as Automatix's RAIL) that can be integrated with robot and industrial manufacturing languages are now beginning to appear and will become commonplace within 5 years.
- Generic representations for amorphous objects (such as trees) have been experimentally utilized and should become generally available within 5 years.

E. Knowledge Acquisition

- Strategies for indexing into a large database of models should be available within the next 2 to 5 years.
- "Training by being told" will supplement "training by example" as computer graphics techniques and vision programming languages become more common.

F. Sensing

- An important area of development is 3D sensing. Several current industrial vision systems are already employing structured light for 3D sensing. A number of new innovative techniques in this area are expected to appear in the next 5 years.
- More active vision sensors such as lidar are now being explored, but are unlikely to find substantial industrial application until the last half of this decade.

- A number of other innovative techniques in 3D sensing are now being developed. Among these are the use of multiple light sources, multiple views, and shape from motion. Some of these techniques may see commercial application within the next two years.

- Kruger and Thompson (1981, p. 33) observe that "By taking several views from particular positions and with carefully controlled illumination, it is possible to separate and independently measure the different surface properties." Industrial vision systems for inspection that use this technique will probably appear within the next several years.

- It is anticipated that within two years solid-state cameras and convolvers will become available that will make stereo machine vision a reality.

G. Industrial Vision Systems

- We will see increased use of advanced vision techniques in industrial vision systems, including gray scale imagery.

- We are now observing a shortening time lag between research advances and their applications in industry. It is anticipated that in the future this lag may be as little as one to two years.

- Advanced electronics hardware at reduced cost is increasing the capabilities and speed of industrial vision, while simultaneously reducing costs.

- Because of low start-up costs and the importance of vision to industrial and other applications, new companies and organizations are rapidly entering the vision field.

- It has been estimated that more than 200 companies are now playing a role in the vision field. A shakeout appears likely as the field settles down, but innovation will continue to encourage new entrants.

- It is anticipated that special lighting and active sensing will play an increasing role in industrial vision.

- Better human/machine interfaces simplifying user reprogramming are now appearing and will become dominant in sophisticated applications within 5 years.

- Common programming languages and improved interface standards will within the next 3 to 10 years enable

easier integration of vision to robots and into the industrial environment.

H. Future Applications

- It is anticipated that about one quarter of all industrial robots will be equipped with some form of vision system by 1990.

- Arden (1980, p. 487) observes that "Increasingly, computer-vision techniques are being applied to real-world problems. This is particularly true of device assembly, circuit board layout, and inspection in the field of industrial automation. Although much of the work is still going on, several convincing demonstration programs have been written, and it is expected that computer vision will soon begin to have a significant impact in industry. At the same time, the computer-vision approach will increasingly be applied to the analysis of images by computer, areas which up to now have been the domain of researchers in pattern recognition--for example, the analysis of handwriting, photomicrographs and radiographs, and satellite imagery."

- It is likely that in the order of 90% of all industrial inspection activities requiring vision will be done with computer vision systems within the next decade.

- New vision system applications in a wide variety of areas, as yet unexplored, will begin to appear within this decade. An example of such a system might be visual traffic monitors at intersections that could perceive cars, pedestrians, etc., in motion, and control the flow of traffic accordingly.

- Computer vision will play a large role in future military applications. The Defense Mapping Agency intends to achieve fully automated production for mapping, charting and geodesy by 1995, utilizing "expert system"-guided computer vision facilities. Other future computer vision military applications include autonomous navigation and guidance for vehicles and missiles, target detection, the interpretation of aerial images for general surveilance purposes and for local battlefield surveilance. Computer vision will also play a large role in future battlefield robots.

- Table X gives Binford's (1982) forecast for computer vision system applications.

Table X*

Example Future Applications for Computer Vision Systems

Short term (1-2 years)

- Industrial Vision Systems
- Cartography: Semi-Automated stereo for terrain mapping

Mid term (2-3 years)

- Cartography: Semi-Automated stereo mapping of complex cultural sites
- Photointerpretation — Monitoring of selected objects in restricted situations

Long term (3-5 years)

- 3D Systems for:

  -warehousing
  -handling unoriented parts
  -inspection of non-laminar parts

- Cartography - automatic feature classification
- Photointerpretation - Automatic classification of a greater variety of objects with greater detail

Greater than 5 years

- Robotic operations in hazardous environments
- Autonomous navigation
- Vehicle Guidance
- Medical image analysis
- Aids to handicapped

More than a decade

- Home robots
- General robotic activities
- Observations of extra-terrestrial bodies

*Based on Binford (1982)

I. Conclusion

In conclusion, the amount of activity and the many researchers in the computer vision field suggest that within the next 5 to 10 years, we should see some startling advances in practical computer vision, though the availability of practical general vision systems still remains a long way off.

REFERENCES

The following abbreviations for some journals and conference proceedings are used:

AI — Artificial Intelligence

CGIP — Computer Graphics and Image Processing

T-PAMI — IEEE Transactions on Pattern Analysis and Machine Intelligence

1AAAI — Firest Annual National Conference on Artificial Intelligence, The American Association for Artificial Intelligence, Stanford University, August 1980

AAAI-82 — Proceedings of the National Conference on Artificial Intelligence, Univ. of Pittsburg, August 1982

4ICPR — Fourth International Joint Conference on Pattern Recognition, Tokyo, November 1978

3IJCAI — Third International Joint Conference on Artificial Intelligence, Stanford University, August 1973

5IJCAI — Fifth International Joint Conference on Artificial Intelligence, Cambridge, Mass., August 1977

6IJCAI — Sixth International Joint Conference on Artificial Intelligence, Tokyo, August 1979

IJCAI-81 — Seventh International Joint Conference on Artificial Intelligence, August 1981

References

Albus, J., Kent, E., Nashman, M., Mansbach, P., and Palombo, L., "A 6-D Vision System," Proceedings of SPIE Technical Symposium East '82, Arlington, VA, May 3-7, 1982.

Aggarwal, J. K., Duda, R. O., and Rosenfeld, A. (eds.). Computer Methods in Image Analysis, IEEE Press & Wiley, 1977.

Agin, G. J. and Binford, T. O. "Computer Descriptions of Curved Objects," 3IJCAI, 1973, pp. 629-640.

Agin, G. J., "Computer Vision Systems for Industrial Inspection and Assembly," Computer, May 1980.

Arden, B. W., (eds.), What Can Be Automated? Cambridge: M.I.T., 1980, pp. 482-487.

Baird, M. L., "Sight-1: A Computer Vision System for Automated IC Chip Manufacture," IEEE Trans. Sys. Man Cybern., vol. SMC-8, 1978.

Ballard, D. H., Brown, C. M., and Feldman, J. A., "An Approach to Knowledge-Directed Image Analysis," in Hanson and Riseman 1978a, pp. 271-281.

Ballard, D. H. and Brown, C. M., Computer Vision, Englewood Cliffs: Prentice Hall, 1982.

Barrow, H. G. and Tenenbaum, J. M., "MSYS: A System For Reasoning About Scenes", SRI AI Center, Tech. Note 121, March 1976.

Barrow, H. G. and Tenenbaum, J. M., "Recovering Intrinsic Scene Characteristics from Images," in Hanson and Riseman, 1978a, pp. 3-26.

Barrow, H. G., "Artificial Intelligence: State-of-the-Art". SRI International, Menlo Park, CA, Tech. Note 198, October 1979.

Barrow, H. G. and Tenenbaum, J. M. "Interpreting Line Drawings as Three-Dimensional Surfaces," IAAAI, 1980, pp. 11-14.

Barrow, H. G. and Tenenbaum, J. M., "Computational Vision", Proceedings of the IEEE, Vol. 69, No. 5, May 1981, pp. 572-595.

Binford, T. O. Inferring Surfaces from Images, Artificial Intelligence, Vol. 17(1-3), 1981, pp. 205-244.

Binford, T. O., "Survey of Model-based Image Analysis Systems," Robotics Research, Vol. 1, No. 1, Spring 1982.

Bolles, R. C. "Verification Vision Within a Programmable Assembly System," AIM-295, STAN-CS-77-591, Computer Science Dept., Stanford University, 1976.

Bolles, R. C., "Robust Feature Matching Through Maximal Cliques," SPIE, Vol. 182, Imaging Applications for Automated Industrial Inspection and Assembly, 1979.

Brady, M., "Computational Approaches to Image Understanding," M.I.T. A.I. Memo No.653, October 1981A. (Also in Computing Surveys, Vol. 14, No. 1, Mar. 1982, pp. 3-71).

Brady, M., "The Changing Shape of Computer Vision," Artificial Intelligence, 17(1-3), 1981B, pp. 1-15.

Brooks, R., Greiner, R. and Binford, T. O., The ACRONYM Model-Based Vision System, Proc. Int. Jt. Conf. Artificial Intelligence 1979, 6, 105-113.

Brooks, R., "Symbolic Reasoning Among 3-D Models and 2-D Images," AI 17, 1981, pp. 285-348.

Brooks, T. L. "Supervisory Manipulation Based on the Concepts of Absolute vs. Relative and Fixed vs. Moving Tasks," Proceedings of the International Computer Technology Conference, sponsored by ASME, San Francisco, August 1980, pp. 185-196.

Chin, R. T., "Automated Visual Inspection Techniques and Applications: A Bibliography," Pattern Recognition, Vol. 15 (4)., 1982, pp. 343-357.

Chin, R., Harlow, C. A., Dwyer, S. J., "Automated Inspection Techniques," in Proc. Assembly IV Conf. Exposition (Detroit, MI), Nov. 1977.

Clowes, M. B., "On Seeing Things," 1971, AI 2, pp. 79-116.

Cohen, P. R., and Feigenbaum, E. A., Chap. XIII, Vision, The Handbook of Artificial Intelligence, Vol. II, Los Altos, CA: Kaufmann, 1982, pp. 125-321.

Draper, S. W., "The Use of Gradient and Dual Space in Line-Drawing Interpretation," AI 17, Aug. 81, pp. 461-508.

Duda, R. O. and Hart, P. E., Pattern Recognition and Scene Analysis, Wiley, 1973.

Eberlein, R. B., "An Iterative Gradient Edge Detection Algorithm," CGIP 5, 1976, pp. 245-253.

Faugeras, O., Price, K., "Semantic Description of Aerial Images Using Stochastic Labelling", Proc. ARPA Image Understanding Workshop, Univ. of Md., April 1980, p.89.

Fennema, C. L. and Thompson, W. B., "Velocity Determination in Scenes Containing Several Moving Objects," CGIP 9, 1979, pp. 301-315.

Fischler, M. A., "On the Representation of of Natural Scenes," in Computer Vision Systems, Hanson and Riseman, (Ed.), (1978a), pp. 47-51.

Fischler, M. A., "Computational Structures for Machine Perception," in Advanced Computer Concepts, J. C. Solinsky (ed.), La Jolla Inst., 1981, pp. 47-55.

Fischler, M. A., and Bolles, R. C., "Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography," CACM, Vol. 24(6), June 1981, pp. 381-395.

Fischler, M. A., Tenenbaum, J. M., and Wolf, H. C., "Detection of roads and linear structures in low-resolution aerial imagery using a multisource knowledge integration technique," CGIP, Vol. 15(3), March 1981, pp. 201-223.

Fischler, M. A., Barnard, S. T., Bolles, R. C., Lowry, M., Quam, L., Smith, G., and Witkin, A., "Modeling and Using Physical Constraints in Scene Analysis," AAAI-82, Aug. 1982, pp. 30-35.

Garvey, T. D., "Perceptual Strategies for Purposive Vision", SRI AI Center Tech. Note 117, 1976.

Jarvis, J. F., "Automated Visual Inspection of Printed Wiring Boards by Local Pattern Matching," IEEE Trans. Patt. Anal. Machine Intelligence, vol. 2, pp. 77-82, Jan. 1980.

Gennery, D. B., "Modelling the Environment of an Exploring Vehicle by Means of Stereo Vision," AIM-339, STAN-CS-80-805, Computer Science Dept., Stanford University, 1980.

Gennery, et al., "Computer Vision", JPL Publ. 81-92, Nov. 1, 1981.

Gevarter, W. B., "A Wiring Diagram of the Human Brain as a Model for Artificial Intelligence," Proc. of the IEEE Inter. Conf. on Cyb. and Society, Wash, D.C., Sept. 1977, pp. 694-698.

Gevarter, W. B., An Overview of Artificial Intelligence and Robotics, Vol. II - Robotics, NBSIR 82-2479, National Bureau of Standards, Wash., D.C., March 1982A.

Gevarter, W. B., An Overview of Expert Systems, NBSIR 82-2505, National Bureau of Standards, Wash., D.C., May 1982B.

Gilbert, A. L. Giles, M. K., Flachs, G. M. Rogers, R. B., and U.Y.H., "A Real-Time Video Tracking System," T-PAMI 2, 1980, pp. 47-56.

Griffin, M. D., Cunningham, R. T., and Eskenazi, R., "Vision-Based Guidance of an Automated Roving Vehicle," paper 78-1294, Proceedings AIAA Guidance and Control Conference, Palo Alto, August 1978.

Hanson, A. J. and Fischler, M. A., "The DARPA/DMA Image Understanding Testbed, Proceedings of the Image Understanding Workshop, Stanford U., Sept. 1982, pp. 342-351.

Hanson, A. R. and Riseman, E. M. (eds.) Computer Vision Systems, New York: Academic Press, 1978a.

Hanson, A. R. and Riseman, E. M. (1978b), "Segmentation of Natural Scenes," in Hanson and Riseman (1978a), pp. 129-163.

Hanson, A. R. and Riseman, E. M., (1978c), "VISIONS: A Computer System for Interpreting Scenes," in Hanson and Riseman (1978a), pp. 303-333.

Herman, M., Kanade, T., and Kuroe, S., "Incremental Acquisition of a Three-dimensional Scene Model from Images," Proc. of the DARPA I.U. Workshop, Stanford U., Sept. 1982, pp. 179-192.

Hiatt, B., "Toward Machines that See," Mosaic, Nov/Dec 1981, pp. 2-8.

Hirzinger, G. and Snyder, W., "Analysis of Time-Varying Imagery for Tracking Moving Objects," Proceedings of the International Computer Technology Conference, sponsored by ASME, San Francisco, August 1980, pp. 26-33.

Holland, S. W., Rossol, L. and Ward, M. R., "Consight-I: A Vision Controlled Robot System for Transferring Parts from Belt Conveyors," Computer Vision and Sensor-Based Robots, G. G. Dodd and L. Rossol, Eds. New York: Plenum Press, 1979, pp. 81-100.

Horn, B. K. P., "Artificial Intelligence and the Science of Image Understanding," in Computer Vision and Sensor-Based Robots, G. G. Dodd and L. Rossol (Eds.), N.Y.: Plenum Press, 1979, pp. 69-77.

Horn, B. K. P. and Schunck, B. G. "Determining Optical Flow," Artificial Intelligence, 1981, 17(1-3).

Hsieh, Y. Y. and Fu, K. S., "A Method for Automatic IC Chip Alignment and Wire Bonding," in Proc. IEEE Computer Soc. Conf. Pattern Recognition and Image Processing, (Chicago, IL), 1979.

Hsieh, Y. Y. and Fu, K. S., "An Automatic Visual Inspection System for Integrated Circuit Chips," Comput. Graph Image Processing, vol. 14, 1980, pp. 293-343.

Huffman, D. A., "Impossible Objects as Nonsense Sentences," in Machine Intelligence 6, B. Meltzer and D. Michie (eds.), Halsted Press, 1971, pp. 295-323. (Also in Aggarwal et al. (1977), pp. 338-366,)

Ikeuchi, K. and Horn, B. K. P., "Numerical Shape from Shading and Occluding Boundaries," Artificial Intelligence, 1981, 17.

Jarvis, J. F., "A method for automating the visual inspection of printed wiring boards, in Proc. SITEL-LILG Seminar Pattern Recognition (Univ. Leige, Sart-Telman, Belgium), Nov. 1977.

Kanade, T., "Model Representations and Control Structures in Image Understanding," Proc. of IJCAI-77, Cambridge, Aug. 1977, 1074-1082

Kanade, T., "Recovery of The Three-Dimensional Shape of An Object from a Single View," AI 17, Aug. 81, pp. 409-460.

Kashioka, S., Fjiri, M., and Sakamoto, Y., "A Transistor Wire Bonding System Utilizing Multiple Local Pattern Machine," IEEE Trans. Syst. Mar. Cybern., vol. SMC-6, 1976.

Kelly, M.D., "Edge Detection in Pictures by Computer Using Planning," in Machine Intelligence6, B. Meltzer and D. Michie (eds.), Halsted Press, 1971, pp. 397-409. (Also in Aggarwal et al. (1977), pp. 228-244.)

Krakauer, L. and Pavlidis, T., "Visual Printed Wiring Board Fault Detection by a Geometrical Method," in Proc. COMSAC (Chicago, IL). Nov. 1979, pp. 260-265.

Kruger, R. P. and Thompson, W. B., "A Technical and Economic Assessment of Computer Vision for Industrial Inspection and Robotic Assembly," Proceedings of the IEEE, Vol. 69, No. 12, Dec. 1981, pp. 1524-1538.

Landgrebe, D. A., "Analysis Technology for Land Remote Sensing," Proc. of the IEEE, Vol. 69, No. 5, May 1981, pp. 628-642.

Levine, M., "A Knowledge-Based Computer Vision System", in Computer Vision Systems, Hanson, A., Riseman, E., eds., Academic Press, New York, 1978.

Lowe, D. G., and Binford, T. O., "The Interpretation of Geometric Structure from Image Boundaries," Proc. Image Understanding Workshop, ed., Lee Baumann S., 39-46, 1981.

MacVicar-Whelan, P. J., and Binford, T. O., "Line Finding With Subpixel Precision," Proc. Image Understanding Workshop, ed., Lee Baumann S, 26-31, 1981.

Marr, D., "Representing Visual Information," in Computer Vision Systems, A. Hanson and E. M. Riseman, Eds.; New York: Academic Press, 1978, pp. 61-80.

Marr, D. C., Vision, San Francisco: W. H. Freeman, 1982.

Marr, D. and Nishihara, H., "Visual Information Processing: Artificial intelligence and the Sensorium of Sight," Technology Review, October 1978, pp. 28-47.

Marr, D. and Hildreth, E. "Theory of Edge Detection," AI Memo No. 518, Artificial Intelligence Laboratory, Mass. Institute of Technology, April 1979.

Martelli, A. "An Application of Heuristic Search Methods to Edge and Contour Detection," CACM 19, 1976, pp. 78-83. (Also in Aggarwal et al. (1977), pp. 217-227.)

Minsky, M. L., "A Framework for Representing Knowledge," In. P. H. Winston (Ed), The Psychology of Computer Vision, New York: McGraw-Hill, 1975, pp. 211-277.

Nagao, M., Matsuyama, T., Ikeda, Y., "Region Extraction and Shape Analysis of Aerial Photographs", Proc 4ICPR, 1978, p. 620.

Nagao, M., Matsuyama, T., A Structural Analysis of Complex Aerial Photographs, Plenum, 1980.

Nagel, R. N., Vanderbrug, G. J., Albus, J. S., and Lowenfeld, E., "Experments in Part Acquisition Using Robot Vision," SME Tech. Paper MS79-784, 1979.

Nevatia, R. and Babu, K. R., "Linear Feature Extraction and Description," 6IJCAI, 1979, pp. 639-641.

Nevatia, R., Machine Perception, Englewood Cliffs, NJ: Prentice-Hall, 1982.

Nitzan, D., Rosen, C., Agin, G., Bolles, R., Gleason, G., Hill, J., McGhie, D., Prajoux, R., Park, W., and Sword, A., "Machine Intelligence Research Applied to Industrial Automation" 9th Report, SRI International, August, 1979.

Nudd, G. R., "Image Understanding Architectures," Proc. of the National Computer Conf., 1980, pp. 377-386.

Ohta, Y., "A Region-Oriented Image-Analysis System by Computer", Thesis Dept. of Information Science, Kyoto University, 1980.

Parma, C. C., Hanson, A. M., Riseman, E. M., "Experiments in Schema-Driven Interpretation of a Natural Scene", Univ. of Mass. COINS Tech. Rept. 80-10, 1980.

Perkins, W. A. "Model-Based Vision System for Industrial Parts," IEEE, Trans. Comput., vol. C-27, 1978.

Pinkney, H. F. L., "Theory and Development of an On-Line 30 Hz Video Photogrammetry System for Real-Time 3-Dimensional Control," Proceedings of the ISP Symposium on Photogrammetry for Industry, Stockholm, August, 1978.

Pratt, W. K., Digital Image Processing, New York: Wiley, 1978, pp. 568-587.

Prewitt, J.M.S., "Object Enhancement and Extraction," in Picture Processing and Psychopictorics (B. S. Lipkin and A. Rosenfeld, eds.) New York: Academic Press, 1970, pp. 75-149.

Reddy, R. and Newell, A., "Image Understanding: Potential Research Approaches." ARPA Image Understanding Workshop, Washington, DC, 1975.

Roach, J. W. and Aggarwal, J. K., "Computer Tracking of Objects Moving in Space," T-PAMI 1, 1979, pp. 127-135.

Roberts, L. G., "Machine Perception of Three-Dimensional Solids," in Optical and Electroptical Information Processing, J. T. Tippitt et al., Eds., Cambridge, MA, MIT Press, 1965.

Rosenfeld, A., "Image Pattern Recognition", Proceedings of the IEEE, Vol. 69, No. 5, May 1981, pp. 596-605.

Rosenfeld, A., "Picture Processing, 1981, "Computer Vision Lab Rept. TR-1134, U of MD, College Park, Jan. 1982. (Also in CGIP, May 1982.)

Rubin, S., "The ARGOS Image Understanding System", Proc ARPA IU Workshop, Nov. 1978, also "The ARGOS Image Understanding System", Thesis, Carnegie-Mellon University, 1978.

Sarris, V., "The Development of Robot Vision", Charles River Associates, Boston, Mass., 1982.

Saund, E., Gennery, D. B., and Cunningham, R. T. (1981), "Visual Tracking in Stereo," Joint Automatic Control Conference, sponsored by ASME, University of Virginia, June 1981.

Shapiro, L. G., Moriarty, J. D., Mulgoankar, P. G., and Haralick, R. M., "Sticks, Plates, and Blobs: A Three-Dimensional Object Representation for Scene Analysis," IAAAI, 1980 pp. 28-30.

Shirai, Y., "Analyzing Intensity Arrays Using Knowledge About Scenes," in Winston (1975), pp. 93-114.

Shirai, Y., "Recognition of Real-World Objects Using Edge Cue," in Hanson and Riseman (1978a), pp. 353-362.

Stevens, K. A., "The Visual Interpretation of Surface Contours," AI 17, August 81, pp. 47-74.

Tenenbaum, et al., "Prospects for Industrial Vision", in Computer Vision and Sensor-Based Robots", G. G. Dodd and L. Rossol (Eds.), NY: Plenum Press, 1979, pp. 239-259.

Tsugawa, S., Yatabe, T., Hirose, T., and Matsumoto, S., "An Automobile with Artificial Intelligence," IJCAI, 1979, pp. 893-895.

Ullman, S., The Interpretation of Visual Motion, Cambridge, Mass: MIT Press, 1979.

Vamos, T., Bathor, M., and Mero, L., "A Knowledge-Based Interactive Robot-Vision System," 6IJCAI, 1979, pp. 920-922.

Waltz, D. L., "Generating Semantic Descriptions from Drawings of Scenes with Shadows", M.I.T. Tech. Rept. AI-TR-271, Cambridge, MA, Nov. 1972.

Waltz, D., "Understanding Line Drawings of Scenes with Shadows," in Winston (1975), pp. 19-91.

Wesley, M. A., Lozano-Perez, T., Liberman, L. I., Lavin, M. A., and Grossman, D. D., "A Geometric Modeling System for Automated Mechanical Assembly," IBM Journal of Research and Development 24, 1980, pp. 64-74.

Winston, P. H. (ed.), The Psychology of Computer Vision, McGraw-Hill, 1975.

Woodham, R. J., "Analyzing Curved Surfaces Using Reflectance Map Techniques," in Artificial Intelligence: An MIT Perspective, P. H. Winston and R. H. Brown (eds.), MIT Press, 1979, pp. 161-182.

Woodham, R. J., "Analyzing Images of Curved Surfaces," AI 17, August 81, pp. 117-140.

Yachida, M. and Tsuji, S., "A Versatile Machine Vision System for Complex Industrial Parts," IEEE Trans. Comput., vol. C-26, 1977.

APPENDICES

APPENDIX A*

LOW LEVEL FEATURES

The scene to be analyzed is usually sensed by a digital camera or other similar device, the output of which is normally a digitized image having an array of brightness values. For some purposes these brightness values can be operated upon directly to obtain desired information about the scene, but it is usual to extract low level features for further computer processing. The following sections describe the low level features usually considered for extraction.

A. Pixels (Picture Elements)

Pixels are the individual elements in a digitized array. They usually represent brightness and perhaps color in a projection from a three dimensional scene, but could also represent distance in a range image.

B. Texture

Texture is a local variation in pixel values that repeats in a regular or random way across a portion of an image or object. Texture can sometimes be used to identify the object being sensed, or it can be used for approximating range and surface orientation in a known object. However it can also be a noise source in processing the image.

C. Regions

A region is a set of connected pixels that show a common property such as average gray level, color or texture in an image.

*This appendix is based largely on Gennery et al. (1981).

D. Edges and Lines

An edge is a step in pixel values (exceeding some threshold) between two regions of relatively uniform values. A line is defined as a thin region of roughly uniform pixel values between two regions of different but roughly equal pixel values. Line representations are extracted from edges.

E. Corners

A corner is an abrupt change in direction of a curve. Corners are useful in data compression approaches to representing straight edges, and as points for feature matching.

APPENDIX B

EXTRACTING EDGES AND AREAS

A natural first step in analyzing a scene is to convert it into a sketch, that is, find the edges that separate regions of differing brightnesses. Edges correspond to abrupt changes in brightness. Such changes can be identified as places where the first derivative of the brightness is suddenly high or the second derivative is zero (see Figure 8). There are various schemes for doing this, all in some way related to taking brightness differences between adjacent points.

A. Extracting Edges

The basic methods for extracting edge and line elements from images are*:

1. Linear Matched Filtering:

Successively convolve** image windows with a template of the desired feature and seek the maximum value.

2. Non-linear Filtering:

Convolve windows in the image with a local operator (weighting function that approximates first or second derivatives by first or second differences). Examples of operators for doing this are shown in Table I. In general, each point in the image is convolved with directional operators in as many directions as

*This section is based largely on methods described in Rosenfeld (1981, p. 601), Gennery, et al., (1981, pp. 2-8 to 2-14), and Brady (1981A). Additional material can be found in Ballard and Brown (1982), Binford (1981), and Nevatia (1982).

**Convolve means superimposing a nxn operator over a nxn pixel area(window) in the image, multiplying corresponding points together and summing the result.

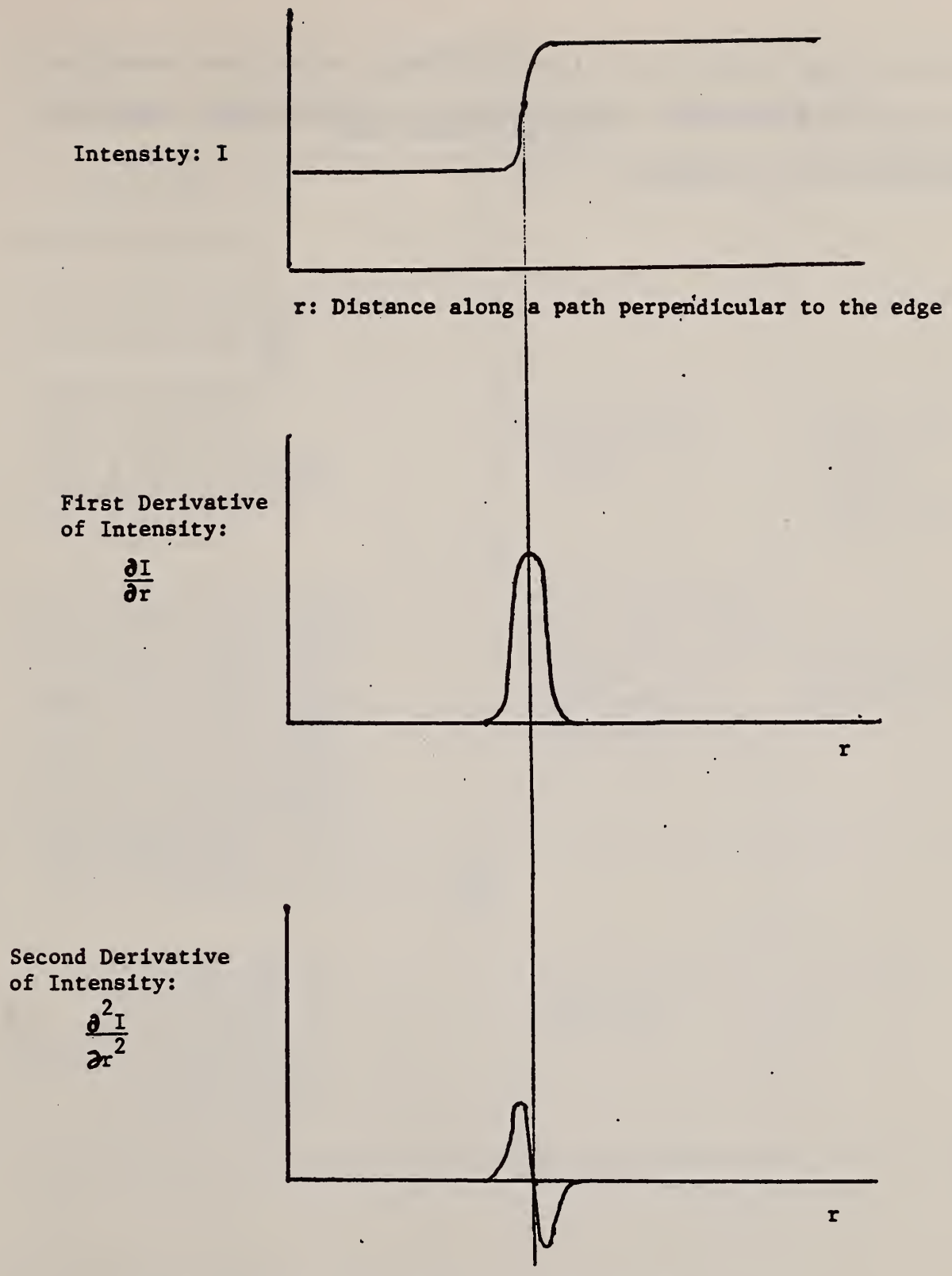


Figure 8 Intensity Variations at Step Edges.

needed. The resultant outputs at each point are combined to determine the gradient vector (the orientation and magnitude of the intensity changes).

Table I

Examples of Non-Linear Filtering* for Extracting Edge and Line Elements

Approach	Edge Criteria	Remarks
I. Edge Operators Detect first derivative of brightness: $\frac{\partial f}{\partial x}$ Sobel Operators Edge Mask -1 0 1 -2 0 2 -1 0 1	pixels which yield max. values	operators tuned for limited range of operation
Nevatia and Babu Operators -100 -100 0 100 100 -100 -100 0 100 100 -100 -100 0 100 100 -100 -100 0 100 100 -100 -100 0 100 100	For each pixel, find angle operators that yield maximum value. Then thin, threshold and link (line fit).	similar operators used for each 30° angle.
II. Bar Operators Detect second derivative: $\frac{\partial^2 f}{\partial x^2}$ Bar Mask -1 2 -1 -1 2 -1 -1 2 -1 -1 2 -1	look for zero crossing	sensative to noise

*Convolving a image window (about a pixel) with operators such as those indicated. Operators shown are for finding vertical lines.

3. Local Thresholding

Apply local thresholding and discard responses that do not lie on borders (between upper and lower threshold regions) and link responses that do.

4. Surface Fitting - The Hueckel Operator

Fit a surface to neighborhood of each pixel and compute maximum gradient of the surface. Consider as edge points those pixels having surface maximum gradients above a selected threshold value. This approach was first devised by Prewitt (1970). The Hueckel Operator is a popular method for doing this.

5. Rotationally Insensitive Operators:

The Laplacian Operator ($\nabla^2 I = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}$, related to the magnitude of the derivative of the intensity gradient) is insensitive to the direction of a line and yields edge elements at pixel points where the Laplacian is zero. Thus discrete approximations to the Laplacian have proved useful in line finding.

6. Line Following:

Shirai (1975) devised a line following method that used a pair of parameters that varied according to how continuously and smoothly elements were found. These parameters determined thresholds for accepting a new element according to how close it was to the linear continuation of the current line being tracked.

7. Global Methods

- Martelli (1976) devised a global hueristic search that operates directly on the brightness values. A cost function is optimized depending on the curvature of the candidate

line and the degree to which the candidate line succeeds in dividing the image into regions of different brightnesses.

- Kelly (1971) used a hierarchical refinement approach, first finding lines in a coarse image and using the results to guide line finding in a higher resolution image.
- Eberlein (1976) utilized a relaxation approach for linking edges found by a local detector, depending on how the edges agreed with their local neighbors. This was a parallel method that merged the elements into a continuous line.
- The Hough Transform (Duda and Hart, 1973), is a global parallel method for finding straight or curved lines. For a straight line, using results from local edge detectors, the perpendicular distance (p) from the line element to the origin and the angle (θ) of the normal to the line is determined and mapped into (p,θ) space. Peak clusters in (p,θ) space are considered to be straight lines.
- Fischler, Tenenbaum and Wolf (1981) describe a new paradigm for detecting and precisely deliniating roads and similar "line-like" structures appearing in low-resolution aerial imagery: The approach combines "local information from multiple, and possibly incommensurate, sources, including various line and edge detection operators, map knowledge about the likely path of roads through an image, and generic knowledge about roads (e.g. connectivity, curvature, and width constraints). The final interpretation of the scene is achieved by using either a graph search or dynamic programming techniques to optimize a global figure of merit."

B. Edge Finding Variations

There are many approaches which can be considered to be variations, combinations and extensions of the basic approaches to edge finding considered in section A.

For example, Marr and Hildreth (1979) utilized the fact that different edges are found depending upon the size of the edge masks. They also observed that bar masks seem to give more reliable information than edge masks. They used bar masks of different panel widths and combined their outputs to reduce effects of noise and to compute the fuzziness of an edge. They extended this method based on their observations that intensity changes are localized in space and in (spatial) frequency. They note that using a Gaussian filter* optimized localization in both domains simultaneously. They thus convolved the original image with the Laplacian of the Gaussian smoothing filter for each spatial frequency used. Edges were considered to occur where zero crossings from several spatial frequency channels concurred.

C. Linking Edge Elements, and Thinning Resultant Lines

Due to imperfections in edge element finding techniques, situations where edges are poorly defined and noise in the image, the primal sketch will usually consist of discontinuous and somewhat scattered edge elements. Various schemes (heuristics) exist to connect these edge elements together to form lines.

*An averaging procedure about a pixel in which the influence of neighboring pixels fall off with distance, according to a Gaussian distribution.

Long edges or lines can be found either by using an edge detector (as discussed in the previous section) and linking the resultant edge elements into a long smooth curve (filling in gaps and ignoring stray elements), or by a procedure which accomplishes a similar result by operating directly on the image data. In either case, if the algorithm operates sequentially by proceeding along the curve as it links edge elements or pixels, it often is called a line follower (or tracker), edge follower, or curve follower. However algorithms have also been devised that operate on an effectively parallel or gestalt basis.

Eberlein's (1976) relaxation method yields a thin line naturally upon convergence. Nevatia and Babu's (1979) approach accepts as edge portions those candidate edge elements found that have a maximal gradient value compared to adjacent pixels with a similar gradient orientation.

When deriving curves from edge data, it is often desirable to thin the resulting contours. Thinning methods reduce the contours to a single-pixel width by discarding redundant edges while maintaining the continuity of the contours. Some methods such as Eberlein's or Nevatia and Babu's include thinning as an inherent part of their operation.

D. Remarks on Edge Finding

Binford (1981) states that it is important to distinguish between detection of an intensity change and its subsequent localization. Thus, he considers the zero crossing of the second derivative of the intensity good for localization of feature points but not for detection; while the maximum of the first derivative is good for detection, but not for localization. Combining the two effects and using linear interpolation,

MacVicar-Whelan and Binford (1981) report being able to localize edges to sub-pixel accuracy.

Gennery, et al., (1981) state that for poor quality images, the performance of all the various detectors degrade, but in different ways. None can be considered to be the last word in edge detectors.

E. Extracting Regions

Many of the edge finding approaches are designed to perform best when the edges can be approximated reasonably by a series of linked straight lines. In natural scenes, this approach can lead to difficulties.

An alternative approach to edge finding is to partition an image into regions of approximately uniform brightness corresponding to surfaces. Unlike edge linking, "region growing" does not require the assumption that the boundaries are straight. Region growing can be accomplished by initially partitioning the image into elementary regions of constant brightness, and then successively merging adjacent regions having sufficiently small brightness differences, until only boundaries with strong contrast remain. The merging can be done somewhat in parallel by computing merge merits for all pairs of adjacent regions, and merging all pairs that have mutually highest merit. Another advantage of region growing over edge finding is that this technique generalizes more readily to characteristics other than brightness, such as texture, color, size and shape, which are important in natural scenes.

The simplest vision systems use a global threshold to obtain a binary image - an approach commonly used in industrial vision

systems. Thresholding can also be local or even dynamic. When histograms* of image pixel intensitites are used, it is usual to dissect the image by thresholding at a value in a valley of the histogram so as to give strong peaks on either side of the threshold value. This region splitting approach can be applied recursively until no more regions can be split. Ohlander et al. (1978) used this approach, computing histograms in each of nine colors and thresholding on the parameter that yielded the best histogram for splitting.

NASA has employed spectral analysis for segmenting regions in LANDSAT imagery (c.f., Landgrebe, 1981).

*Frequency counts of the occurrence of each intensity in an image.

APPENDIX C

SEGMENTATION AND INTERPRETATION

A. The Computer Vision Paradigm

Starting with an image of a scene, the goal of a computer vision system is to identify the objects and their relationships in the scene. To accomplish this, it is customary for the system to segment the image into surfaces or edges associated with the objects, and then use the resulting information, together with domain knowledge, to generate the desired scene description. In this appendix we will review techniques used to do this segmentation.

B. An Early Bottom-up System

A landmark program in machine perception was developed by Roberts (1965) to recognize various three dimensional polyhedral object configurations. Roberts employed an image-dissector camera to look at blocks-world scenes involving blocks, wedges, hexagonal prisms, or objects formed by sticking these together. His program could determine the location, orientation, and dimensions of the objects. The program could demonstrate its "understanding," by displaying a drawing of the scene observed from any desired viewpoint.

Roberts' program first found the places in the images where brightness or shading changed abruptly, corresponding to points on the edges of the object. Then by linking these points, it produced a line drawing of the scene. The line drawing was interpreted by finding triangles, quadrilaterals, and hexagons, which suggested possible objects (triangles suggest wedges, etc.) and eventually accounted for all the lines and junctions as

edges and corners of objects. From the resulting appearance of the object in the image, the program was able to compute its dimensions, location, and orientation.

C. Problems with Bottom-Up Systems

Barrow and Tenenbaum (1981, p. 570) note that a major problem with sequential program organization used by Roberts and many of his successors:

> ...is the inherent unreliability of segmentation. Some surface boundaries may be missed because the contrast across them is low, while shadows, reflections, and markings may introduce extra lines and regions. The interpretation phase, when presented with a corrupted segmentation, may be unable to produce an explanation, and hence cause the entire system to fail.
>
> Partitioning an arbitrary image into regions corresponding to objects or object surfaces is fundamentally impossible without exploiting scene models. First, there is no basis for deciding which image features are significant at the level of objects and which are not. Second, there is no good pictorial criterion for filling in missing features. Third, the very notion of an object is ill defined, being largely determined by convention and experience.

D. Interpretation-Guided Segmentation

Several research teams tried to overcome these problems by integrating the segmentation and interpretation phases. One simple approach used was to try to recognize objects from partial matches obtained using models and then to try to verify the results by attempting to find evidence that supported image features previously missed.

Techniques were also developed for region-based systems. The general approach being:

1. For regions with uniform attributes such as intensity, color or texture, assign sets of possible object

interpretations based on knowledge of possible object surfaces and the contextural contraints associated with assignments in adjacent regions. For instance, a road cannot be surrounded by sky.

2. Merge adjacent regions with comparable interpretations.
3. Reevaluate interpretations based on contextual constraints associated with the new adjacent regions.
4. Continue alternating merging and interpreting until all adjacent regions have disjoint interpretations unviolated by contextual constraints.

Both line-based and region-based interpretation-guided segmentation systems have been devised that have performed well in a variety of complex scene domains. However, the approach is not suitable for a general-purpose vision system as it depends on prior knowledge of expected objects. Unknown objects cannot be recognized or even described. Thus for unknown objects, levels of scene descriptions below the level of complete objects are needed.

E. Use of General World Knowledge to Guide Segmentation

Marr and Nishira (1978) observe that as the primal sketch is typically a large and unwieldy collection of data, the next step is to decode it--traditionally by "...a process called segmentation whose purpose is to divide a primal sketch, or more generally an image, into regions that are meaningful, perhaps as physical objects." It makes sense to use any general knowledge that might help in the interpretation. An example of such

knowledge is information on the physical nature of edges of objects.

Huffman (1971) and Clowes (1971) devised an approach to enable the interpretation of perfect line drawings of polyhedral objects without having to resort to heuristics. They recognized that each line in the picture represented either a convex edge, a concave edge, or an occluding edge in a three-dimensional scene. From this, they constructed a catalog of possible vertices with allowable line labellings. A scene could then be analyzed by starting at one vertex and proceeding through the line drawing performing a tree search, limiting the number of possible line labellings at each step according to the catalog, until a consistent labelling for the entire scene is obtained. Waltz (1975) further extended this technique to include shadows and cracks. His catalog included several thousand possible vertex types. He used a relaxation-type procedure to decide on the correct labelling for each line according to the possibilities in the catalog. The resultant procedure converges rapidly (usually to a unique interpretation) regardless of the complexity of the scene.

A move toward a more general approach to the problem of interpreting feature point segments as lines and edges has recently been made by Binford (1981) and Lowe and Binford (1981). In their scheme, a segment is interpreted as a space curve, and constraints are formulated based on coincidence, and those situations in which a curve corresponds to a true edge or bounding contour.

A comprehensive approach to deriving a physical sketch of a scene from one or more images has been taken by Fischler et al. (1982). They use a priori knowledge of global and extended constraints to guide the segmentation and interpretation process. Their approach involves modeling physically meaningful information such as the imaging process, the scene geometry and elements of the scene content. They utilize knowledge about such factors as the camera model, vanishing points, geometric distortion, ground plane, geometric horizon, skyline, semantic context (urban or rural scene, etc.) physical surface models and edge classification.

APPENDIX D

2-D REPRESENTATION, DESCRIPTION AND RECOGNITION

This appendix presents a number of 2-D representations and descriptions useful for further processing and recognition.

A. Pyramids

A pyramid data structure represents an image at several levels of resolution simultaneously. The base of the pyramid is the original full resolution image, usually assumed to be a n x n square array. The next level of the pyramid is typically formed by partioning the image into non-overlapping 2 by 2 cells and mapping (usually by average gray level) the four pixels in each cell to a single pixel in the next level*. This is repeated, level by level, until the image is compressed into a single pixel at the top level. The usefulness of pyramids lies in being able to extract features at an appropriate level of resolution.

B. Quadtrees

A quadtree representation of a n x n image is obtained in a top-down manner by recursively splitting the image into quadrants, the quadrants into subquadrants, etc. The process continues until all pixels in a quadrant are uniform with respect to some feature (such as gray level). The terminal leaves of a quadtree are uniform regions of varying sizes, thus being a useful first phase in segmenting a image into regions.

C. Statistical Features of a Region

Once an image has been segmented, a description of each region, or blob, can be generated as a list of statistical

*Other partionings and mappings are common.

features. These features typically include perimeter, area, c.g., first and second order moments, color, etc. The individual blob descriptors are linked to form a tree data structure which represents nesting relationships. The parent of any blob in the tree is the adjacent blob which completely surrounds it. Recognition is performed by matching the statistical features with those of stored prototypes. The SRI Vision Module and GM's CONSIGHT use this approach.

D. Boundary Curves

The boundary of a region can be represented by a chain of straight lines and arcs. The resulting compressed boundary descriptions are sometime referred to as "chain codes" or "concurves." Gennery et al. (1981, p. 3-2) note that "The main advantage of the concurve representation is that objects may be recognized on the basis of partial views by matching a subset of the lines and arcs in a model concurve with the image data."

E. Run-Length Encoding

For a binary image, it is possible to segment the image into edges and regions by sequentially scanning the image and recording the edge points (where pixels change from zero to one or vica versa). This process of reducing a binary image to a set of edge points is called run-length encoding, and has been successfully used in the SRI Vision Module and a number of sophisticated commerical vision systems derived from that module.

F. Skeleton Representations and Generalized Ribbons

In this approach, a planar region is represented by a skeleton which consists of the medial line (locus of points

equidistant from the boundaries of the region) and the perpendicular distance from the boundary for each point on the medial line. In some cases, a complex region can be constructed as the union of these generalized ribbons (the 2-D version of generalized cones described in Appendix F).

G. Representation by a Concatenation of Primitive Forms

A region can be built up from a collection of squares, rectangles or other shapes. The "Maximal Block" approach uses a union of squares of various sizes.

H. Relational Graphs

An image that has been segmented into regions can be described in terms of a relational graph, whose nodes represent regions and whose arcs represent properties (such as shape and size) and relations (such as "infront of" and "adjacent to"). Corresponding views of known objects can be similarly represented, and recognition can be achieved by matching the graphs.

I. Recognition

Recognition consists of matching a description derived from an image to a description of a stored model. Recognition can be accomplished by correlation, which for binary data reduces to template matching. A more elaborate approach is statistical pattern classification using features such as described in Section C. Relaxation and syntactic analysis approaches (described elsewhere) have also been used. Fischler and Bolles (1982) suggest "random sample consensus" as a paradigm for selecting the model that provides the best match to the data and for computing the best values of the free parameters.

APPENDIX E

RECOVERY OF INTRINSIC SURFACE CHARACTERISTICS

A. Basic Approach

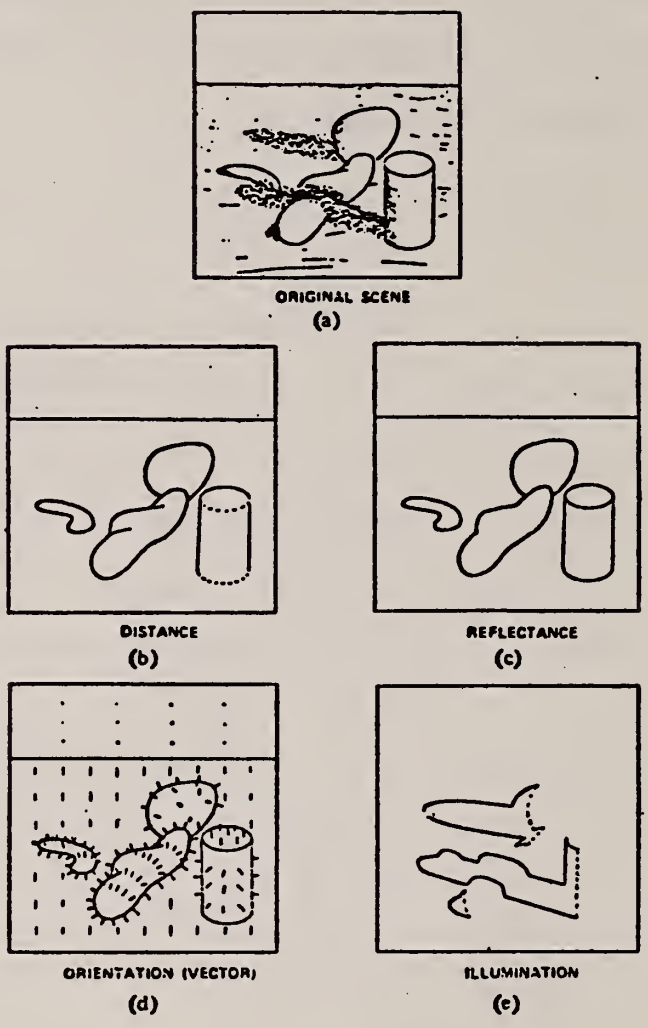
As indicated earlier, it is helpful in many cases to assist in finding 3-D surfaces and volumes for interpretation, to go beyond the 2-D representation of edges and regions to a representation proposed by Marr (1978) of MIT, called the 2.5-D sketch consisting of surface distances and orientations. Such a sketch can be constructed from the surface characteristics which are intrinsic to the scene and are not dependent upon idiosyncracies of viewpoint of the sensor.

Barrow and Tenenbaum (1981, pp. 581-582) indicate that these intrinsic characteristics of surfaces are appropriately represented as a set of arrays in registration with the image array. Each array corresponds to a particular intrinsic characteristic such as surface reflectance, surface orientation, incident illumination and range. Each array contains values for its intrinsic characteristic at the surface element visible at the corresponding point in the sensed image. It also explicitly indicates boundaries due to discontinuties in value or gradient of the characteristic. Such arrays have been referred to as intrinsic images.

Figure 9 is an artist's conception of one possible set of intrinsic images, corresponding to a monochrome image of a simple scene. The images are shown as line drawings, but in fact would contain values at every point. The solid lines represent discontinuities in the scene characteristic; the dashed lines

represent discontinuities in its derivative. The distance image gives the line of sight range from the center of projection to each visible point in the scene. The reflectance image gives the albedo (the ratio of total reflected to total incident illumination) at each point. The orientation image consists of vectors representing the direction of the surface normal at every point. The integrated incident illumination from all sources is given by the illumination image.

Figure 9: A set of intrinsic images derived from a single monochrome intensity image.



Source: Barrow and Tenebaum, 1981, p. 582.

The central problem in recovering the intrinsic characteristics from the image is that the desired information is confounded in the sensory data. The observed light intensity at a single point could result from an infinitude of combinations of illumination, reflectance and orientation. The key to recovery lies in exploiting constraints derived from assumptions about the nature of the scene and the physics of the imaging process. For example, as surfaces are continuous except at boundaries, we can expect surface characteristics (reflection, orientation and range) to also be continuous. Similarly, incident illumination also varies smoothly over a scene except at shadow boundaries.

Barrow and Tenenbaum (1981, p. 589) propose the following four-step model for using interacting constraints in a relaxation type process for simultaneously recovering the primary intrinsic characteristics from a brightness image:

1) find the brightness discontinuities in the input image;
2) determine the physical nature of the discontinuity;
3) assign boundary values for intrinsic characteristics along the edges, based on the physical interpretation;
4) propagate from these boundary values into the interiors of regions, using continuity assumptions.

Many different approaches to recover shape from image characteristics have been explored as represented by the following sections.

B. Shape from Shading

Barrow and Tenenbaum (1978) describe a low level method of estimating relative distance and surface orientation from a

single image. They use heuristics based on the rate of change of brightness across the image.

Ikeuchi and Horn (1981) have formulated a second order differential equation which Horn calls the "image irradiance equation." This equation relates the orientation of the local surface normal of a visible surface, its surface reflectance characteristics, and the lighting, to the intensity value recorded at the corresponding point in the image.

C. Stereoscopic Approach

Gennery et al., (1981, pp. 6-1 to 6-4) describe various stereoscopic approaches to finding range. They observe that the basic stereo approach uses triangulation between two or more views from different positions to determine distance. However, stereo techniques differ in the way in which matching is done between pictures, particularly in the kind of entities that are matched. The two major approaches are area correlation and matching lines of maximum intensity changes (edge-based stereo).

They report (p. 6-3) that, "Scenes of man-made objects often are not highly textured but contain sharp brightness edges at boundaries of objects and at intersections of planar faces. For such scenes, area correlation does not work very well. Instead, it is usually better to detect features in each image and to match these features."

D. Photometric Stereo

In this approach, the light source illuminating the scene is moved to different known locations, and the orientation of the surfaces deduced from the resulting intensity variations (Woodham, 1979, 1981).

E. Shape from Texture

Brady (1981A, p. 88) reports that, "Of the modules which seem to bridge the gap between the Primal Sketch and the Surface Orientation Map, none has received quite as much attention from Psychologists as the computation of surface orientation and depth from texture gradients." Various methods for computing texture gradients are possible and from this orientation can be deduced.

F. Shape from Contour

Barrow and Tenenbaum [1980] have suggested a method for interpreting curved line drawings as three-dimensional surfaces. To interpret a two-dimensional curve, a three-dimensional curve projecting to it is computed that minimizes a combination of variation in curvature and departure from planarity. Other approaches to this problem are given by Draper (1981), Kanade (1981) and Stevens (1981).

G. Shape and Velocity from Motion

Brady (1981A, p. 96) provides a review of efforts to recover shape from motion for the case of rigid bodies. He reports that Ullman (1978) was the first to treat this issue. He considered the problem of establishing a correspondence between the Primal Sketches in two successive image frames. Ullman also studied the problem of computing the structure of a rigid body from the correspondences of a small number of points in a number of views and found that remarkably few of each are required to compute rigid three-dimensional structure.

Brady (1981A, p. 70) defines "optical flow" as the distribution of velocities of apparent movement caused by

smoothly changing brightness patterns. Horn and Schunck (1981) have proposed a method for computing "optical flow" by differentiating the brightness distribution in successive images with respect to time."

APPENDIX F

HIGHER LEVELS OF REPRESENTATION

The basic form for the higher levels of representation is the 3-D model. This is an object-centered representation that describes the object in a convenient way, as in the following examples.

A. Volumetric Models

1. Generalized Cones

Agin and Binford (1973) introduced the concept of generalized cones (also called generalized cylinders). A generalized cone is defined by a space curve, called the spine or axis, and a planar cross section normal to the axis. A "sweeping rule" describes how the cross section changes along the axis. Complicated objects can often be represented by a concatenation of generalized cones.

2. Wire Frame Models

Various investigators have represented 3-D objects by means of wire frame models in which the wires correspond to edges or boundaries of cross sections. Stick figure models are a related representation.

3. Polyhedral Models

Wesley et al., (1980) report on a geometric modeling system developed at IBM to describe complicated mechanical parts. The object is represented by polyhedral primitives which are combined as required by the operations of union, difference and intersection. In the IBM system, objects and assemblies are represented in a graph structure that indicates part-whole

relationships, attachment, constraint, and assembly. Also included are physical properties of objects and positional relationships between objects. The system can determine the appearance of an object for an arbitrary view. This information provides the potential for use by a computer vision recognition system to guide the search for features to match an image to the model.

4. Combining 1D, 2D, and 3D Primitives

Shapiro et al., (1980) describe objects in terms of the primitives: sticks, plates and blobs. Relations are given on how the parts connect, their size, and spatial relationships.

5. Planes and Ellipsoids

Gennery (1980) produced a method for describing 3-D outdoor scenes. The ground surface was approximated by one or more planes or paraboloids, and objects lying on the ground were approximated by ellipsoids.

6. Sets of Prototype Volumes

Efforts in Computer Aided Design and Computer Aided Manufacturing (CAD/CAM) often represent objects by combining a small set of prototype volumes such as spheres, blocks and triangular prisms.

B. Symbolic Descriptions

The various parts of an object in a scene may be represented by graphs in which the nodes are the objects and the arcs are the relations (such as above, to the right of, behind, surrounded by, part of, larger than, etc.) and intrinsic attributes (e.g., small, flat, etc.).

Barrow and Tenenbaum (1981, p. 576) observe that, "Symbolic

models are appropriate for natural objects (e.g., trees) that are better defined in terms of generic characteristics (e.g., larger, green, leafy) than their precise shape."

C. Procedural Models

Rosenfeld (1981, p. 604) defines a procedural model as any process that generates or recognizes images. An important class of such models are grammatical or syntactic models. Pratt (1978, pp. 574-578) discusses such syntatic processes. He observes that syntatic methods have been proved feasible for simple models, but notes that it is not clear yet whether or not these techniques can be extended to general classes of images.

APPENDIX G

HIGHER LEVELS OF INTERPRETATION

Barrow and tenenbaum (1981, pp. 591-593) outline how interpretation might proceed based on intrinsic images. They observe that intrinsic images provide scene information on a point by point basis in a viewer-centered coordinate frame. Higher levels of interpretation, such as object recognition, require a more global representation in a viewpoint-independent coordinate frame. Surfaces and volumes are obvious candidates for representaions following from intrinsic images.

An interpretation-guided segmentation approach based on structural prototypes is a possible mechanism for deriving 3-D surfaces and volumes from intrinsic images.

Once a scene description has been obtained in terms of surface and volume primitives, geometric models can be used to generate similar primitives, which can then be matched by a search process to obtain object recognition and location. It is often convenient to use graph structures for representing scene parts. As scene descriptions are typically fragmented and include many objects, some of which may be occluded, it is necessary to match parts of the scene graph with parts of object graphs. As such subgraph matching can be combinatorially explosive, much work has been done on algorithms to handle such matching in complex scenes.

Barrow and Tenenbaum suggest that perhaps the best way to defeat the combinatorics of search is to decompose object models hierarchically into components. These components can then be

independently matched, and combined and checked for consistancy afterward. Using this approach, the complexity of matching tends to increase additively rather than exponentially.

APPENDIX H

TRACKING

Gennery et al. (1981, pp. 5-1 to 5-3) survey the real-time tracking problem, observing:

> The goal of object tracking is to process sequences of images in real time to describe the motion of one or more objects in a scene. Often real time implies processing every image from a TV camera operating at 30 Hz. In other words, an image is digitized, features are extracted from the image, the object or objects are located in the image, and position and velocity estimates are updated 30 times a second, although in practice slightly slower rates are sometimes used. At the present time, the approaches which achieve real-time operation rely on simplifying assumptions about the nature of the scene, track very few objects in a given scene, and incorporate varying levels of special-purpose hardware designed for the particular tracking algorithm...
>
> Since successive images are only 1/30 second apart in time, the appearance of the object will change very little from image to image. The object can be modelled adaptively as it was last seen by the tracker, with the expectation that a good match between the object model and the features in the current image is available. Furthermore, the location of the object in the image can be predicted very accurately by using the latest available position and velocity estimates coupled with the short elapsed time between images. As a result, the search window need only be large enough to contain the object up to a few pixels uncertainty. This limits the required computation to a manageable level and, more importantly, greatly reduces the probability of a false match occurring...
>
> Real-time implementations typically rely on features which can be computed directly from the image without resorting to actual 3-D measurements of object features.

Table IX summarizes the various approaches surveyed by Gennery et al. It will be observed that a variety of approaches are possible using either area correlation or feature matching. However, no final optimum system has yet been devised.

Table IX

Visual Tracking Approaches

System Developer	Purpose	Approach	Comments
Griffin et al. (1978)	Object tracking for closed-loop guidance of JPL breadboard Mars-Rover vehicle.	Gray level correlation of a window in current successive images of an object. Implemented in software.	Immune to background changes if tracking window confined to target.
Pinkney (1978)	Control of shuttle manipulator for grasping objects tracked.	Uses a single camera to track four man-made markers on object to derive object position and orientation relative to manipulator.	
Brooks (1980)	Supervisory control of a teleoperator manipulator.	Stereo cameras to track markings on an object.	
Nitzen et al. (1979)	Track moving objects for feedback to an industrial robot.	Use SRI vision module.	
Roach & Apparwal (1979)	Track objects in "blocks world"	Blocks are located and matched (using a three level scheme) based on predictions from stored internal representations of blocks discovered in previous images.	Works well in blocks world.
Fennema & Thompson (1979)	Object tracking	"Gradient Intensity Transform Method" Time variations in intensity and the spatial gradient are determined and recorded for each pixel in image. A Hough transform method is used on the intensity variations and gradients to determine object velocity.	Requires smoothing (and therefore accuracy degradation) for procedure to work well.

Table IX (cont.)

Visual Tracking Approaches

System Developer	Purpose	Approach	Comments
Tsugawa et al. (1979)	Detect position of road features to automatically guide a car.	Differentiate analog video signals from two cameras and stereo-match contrast edges.	
Hirzinger & Snyder (1980)	Object tracking	Analog video signal processed by special purpose hardware to detect significant contrast areas, inside a programmable tracking window. The position of the object is considered to be the centroid of the extremes of the contrast points.	This contour-based approach easily fooled in scenes of moderate complexity.
Gilbert et al. (1980)	Real-Time identification and tracking of missiles and aircraft.	Uses four microprocessors as follows: 1. performs histogram analysis of window in image to classify pixels as (1) belonging to target (0) not belonging to target. 2. Sums target pixels horizontally and vertically to identify target. 3. handles image rotation and camera control. 4. evaluates goodness of match at each tracking iteration and adapts system as needed.	Assigns a confidence level to each match and relies on prediction when match is poor.
Saund et al. (1981)	Object tracking	Use feature-matching to an internal model adjusted by a least squares fit.	Rejects extraneous features not predicted by model.

Source: Derived from Gennery et al. (1981, pp. 5-1 to 5-7).

Real-time tracking is important for manipulation; for guidance for applications such as recovering satellites or free-flying payloads; for grasping moving objects (such as parts in an industrial environment); for assembling objects such as machinery or electrical appliances; and for building space structures. It is also important for target acquisition and tracking or for locking onto a feature in situations such as planetary flybys and astronomical or earth observations. And, as to be expected, it is also applicable for vehicle and missile guidance.

APPENDIX I

Additional Tables of

Model-Based Vision Systems

Table III- a

Model-Based Vision Systems

Developer: Ballard, Brown and Feldman (1978), Univ. of Rochester

Purpose: Answering Queries about Images

Sample Domains: Locating ships at docks
Locating ribs in chest x-rays

Approach	Modeling & Representation	Remarks
System is structured in levels (similar to VISIONS) -The model -Sketchmap relating model to image -Images at different levels of resolution Query determines level of detail Queries take the form of user-written executive programs Control involves synthesis of a sketch	Knowledge and model in terms of semantic networks, consisting of nodes and spatial constraints Templates are used to describe shape User models objects in image domain (not 3D domain)	Uses a fixed and known viewpoint User must code an executive match procedure for the particular task domain Special purpose system

Developer: Bolles (1976), SRI

Purpose: Inspection and Visual Control in Repetitive Manufacturing Tasks

Sample Domains: Location of a mechanical part in an automatic assembly work station

Table III- b

Model-Based Vision Systems

VV (Verification Vision System)

Approach	Modeling & Representation	Remarks
Relies on 3D relationships of observables to locate a mechanical part accurately—based on slight deviations from an expected location and orientation 1) The user chooses potential operator/feature pairs 2) The system applies the operator to several sample pictures and gathers statistical information on their effectiveness 3) System ranks operators and predicts cost of accomplishing task 4) System applies operators to task in order of their cost effectiveness, until desired confidence is reached. Uses a generalized least squares algorithm and maximal clique* finding to match the model to identified features *Maximal cliques are maximal matches of portions of graphs of image and model descriptions	Limited 3D models made up of surface point features and their locations in 3D space.	Depends on small correlation windows as features. Thus it is restricted in viewpoint

Developer: Rubin (1978), CMU

Purpose: Identify Objects in Images

Sample Domains: Labelling buildings in a city scene

Table III- c

Model-Based Vision Systems

ARGOS

Approach	Modeling & Representation	Remarks
Has 3D knowledge of positions of the buildings translated into adjacency information to guide the search for labellings of pixels ARGOS does not segment — it labels. It works with pixels or regions. Much depends on spectral labeling. Search is a very local pixel-based "locus" search, a generalization of HARPY's (network speech understanding) "beam search" to 2D photo interpretation.	Internal model is a 3D model of the city. Model used to generate all possible views Stores multiple representations of buildings, in terms of such things as location, texture, color, orientation and gross shape features — all gleaned from training examples	Viewpoint dependent Mostly relies on adjacency relations Not readily generalizable

Developer: Garvey (1976), SRI

Purpose: Locate Known Objects in an Image

Sample Domains: Office environment

Table III- d

Model-Based Vision Systems

Approach	Modeling & Representation	Remarks
System uses simple local features rather than structured shape descriptions Strategy is to: -use windows to acquire image samples which might belong to the object -hypothesize the object from the sample -validate the hypothesis -erect a boundary around the object Coarse to fine strategy: -large objects found first, reducing search area for spatially-related smaller objects Top-down approach	Programmed interactively Objects are shown to the system by outlining them in an image Objects are automatically characterized by conjunction of histograms of local surface attributes such as: hue, orientation, range and height, and relationships between surfaces	Performance rests strongly on having depth data and surface orientation derived from depth Does not use general shape information

Developer: Barrow and Tenenbaum (1976), SRI

Purpose: Identify known objects in a scene

Sample Domains: Typical objects in a room
System also used to drive the IGS interpretation-guided segmentation system

Table III- e

Model-Based Vision Systems

MSYS

Approach	Modeling & Representation	Remarks
An interpretation-guided segmentation system Simulated range data is used to determine 3D locations, orientations of regions and their spatial relations Matcher uses spatial relationships as a strong constraint in matching data from image to model	Models objects in terms of height, orientation and 3D spatial relationships between objects	Relies on a particular viewpoint No shape information is used

Developer: Kanade (1977), CMU

Purpose: Find objects in a scene

Sample Domains: Outdoor scenes with buildings viewed from eye level

Table III- f

Model-Based Vision Systems

Approach	Modeling & Representation	Remarks
Views scene in either intensity or range Tries to match observed patches against modeled patches	Uses a 2-1/2D scene domain Objects are represented as image regions Shape and spatial relations describe the region Objects have multiple representations from multiple viewpoints, but these must be explicitly described by the user	Image oriented Viewpoint is only pseudo-independent

Developer: Nagao (1978, 1980), Kyoto U.

Purpose: Label areas and objects in aerial photographs taken in several spectral bands

Sample Domains: Countryside, Suburban

Table III- g

Model-Based Vision Systems

Approach	Modeling & Representation	Remarks
1. First do edge-preserving smoothing 2. Segment images into regions which are continuous in spectral properties 3. Using histograms and adaptive thresholding in each spectral band, extract cue regions: -large homogeneous areas -elongated regions -shadow and shadow-making regions -vegetation regions -water regions -high contrast texture regions 4. Analyze each cue region by an object detection program specific to region type 5. Feed summary of properties of regions back to subsystems 6. System control tries to resolve conflict labels and to deal with unlabeled regions. The most reliable labels are chosen for a region. If a region can't be labelled, system activates a split and merge process to correct faulty segmentation	Use shadows to give information about height Shadow-making regions are regions adjacent to shadows with a long common boundary Elongated objects include roads, rivers and railroad lines Vegetation areas have small ratios of red to IR Water identified by spectral properties High contrast regions are woods and residential areas Residential regions are those with strong gradients in two orthogonal directions Houses found in candidate residential areas by a sequence of house routines, starting with rectangular-shaped shadow-making regions	Well-crafted system tailored to multi-spectral aerial photographs Segmentation primarily dependent on color Shadow identification not general or reliable Interpretation not general -3D only from shadows -weak use of shape -interpretation suitable for large areas — not human scale objects for which shape is important

Developer: Ohta (1980), Kyoto U.

Table III- h

Purpose: Semantic labelling of regions in color images of outdoor scenes

Model-Based Vision Systems

Sample Domains: Urban scenes with buildings, trees, streets and cars as viewed from ground level

Approach	Modeling & Representation	Remarks
Forms regions by splitting using thresholds from histograms of color parameters. The color parameters chosen are 3 algebraic combinations of red, blue and green $[\frac{r+b+g}{3}, \frac{r-b}{2}, \frac{2g-r-b}{4}]$. Textured regions determined separately based on the Laplacian exceeding threshold in 3x3 windows. A plan is generated by an initial bottom-up coarse region segmentation. A symbolic description of the scene is made by a top-down analysis of the bottom-up interpretation using a production system with knowledge of the world represented as a set of rules. Decisions made by the top-down process cause the bottom-up process to be reactivated to reevaluate the plan. The top-down computation proceeds from a coarse to fine analysis, in a scene phase and an object phase.	Data structure includes regions, boundaries and vertices. Regions are represented by: area, mean intensity of r,b,g, degree of texture, perimeter length, c.g., number of holes, etc. Road model -subobjects: cars, shadows -made of: asphalt, concrete -properties horizontally long, below horizon (car: horizontally long, dark, above road) Sky model -not touching lower edge of image -shining -blue or grey -touching upper edge -linear boundary on lower edge Tree model -heavy texture -made of leaves Building model -subobjects: windows -made of concrete, tile or brick -many holes -many straight lines -linear upper boundary with sky Rules for the bottom-up plan are unary properties of objects and binary relations between objects The world model is a network of objects, materials and concepts (scene schemas).	System does well overall One of a few examples of reasonable performance on scenes of moderate complexity on a set of somewhat different scenes Quality of segmentation is weak Models are weak Approach is ineffective in many situations in which fine details determine object labels.

Developer: Shirai (1978), ETL, Tokyo
Purpose: Recognition and location of common objects from light intensities (grey levels) in an image.
Sample Domain: Generic desk-top objects

Table III- i
Model-Based Vision Systems

Approach	Modeling and Representation	Remarks
Uses an edge-finding process which extracts edges of curved objects. Analysis of the scene starts from the most obvious object. Top-down approach. Recognizes objects using a hierarchy of features. -Find the main feature to get clues for the object. -Find a secondary feature to verify the main feature (and object identity) and to determine the range to the object. -Determine the range and find the other lines of the object. Find secondary small objects after large objects are found	Describes edges by straight lines or elliptic curves Objects are modeled and represented by primary and secondary features Primary / Secondary Lamp: Contour of lamp shade / pair of vertical lines corresponding to trunk; contour of base Bookstand: long vertical lines clustered in a rectangular region / lines connected to main feature Small Objects: Shape and size of contour. (pipe, pen, etc.) / details of shape, and light intensity changes	Edge finder adequate for task Can't deal effectively with texture Use only image models not object models No organization of related edges Top-down approach O.K. with only a few objects

Developer: Levine (1978), McGill U., Canada

Purpose: Develop a Modular Computer Vision System to Experiment with Different Picture Analysis Strategies

Sample Domains: Suburban Outdoor Scenes
Office Scenes

Table III- j
Model-Based Vision Systems

Approach	Modeling & Representation	Remarks
A three level system: First level segments pictures into regions without scene context. Second level has two phases: Local phase matches all stored image templates (of feature vectors) with observed regions, using A* graph search Global optimization phase uses dynamic programming to merge regions and assign labels to them based on model-driven spatial relationships Highest level includes management-type relational data base of image-oriented scenes, and a data-driven production system of what actions to take depending on what appears in the image representation in short term memory (STM). STM contains list of regions and a confidence-ordered list of their interpretations. (It resembles the blackboard of HEARSAY.) Implicit actions are invoked when a region matches an object in a scene model with a confidence level above threshold.	Low level processing is approximately in order of decreasing size using a pyramidal data structure Uses feature vectors throughout Features are stored in three classes according to decreased importance in reducing search time. 1. Includes minimum bounding rectangle and its areas. 2. Intrinsic features: intensity, hue, saturation and its area. 3. Includes six moment invariants as a rough measure of shape, and detailed shape from a set of Fourier coefficients for the outline (used only in final template evaluation).	Segmentation weak—based on a gradient operator System is viewpoint dependent

Developer: Herman, Kanade and Kuroe (1982), CMU

Purpose: To incrementally acquire a 3-D model of a complex urban scene from images.

Sample Domains: Near vertical views of Federal Buildings area of Wash., D.C.

Table III- k
Model-Based Vision Systems

3-D Mosaic Project

Approach	Modeling & Representation	Remarks
Uses stereo analysis to construct partial wire frame models from scene vertices and edges (which have been previously extracted from the images using conventional edge point finding, thinning, and fitting of straight lines). Constructs a structure-graph to represent partial constraints on 3-D structure. From the wire frame descriptions, a surface-based model representing an approximation to the scene is generated using domain-specific knowledge of building shapes - such as flat roofs and vertical sides. Uses a relaxation type process to merge wire frame models generated from different stereo pairs. Modifications, additions or deletions to the structure-graph model are made as new information is found.	Constructs wire frame models from images. Uses a structure-graph representation to model surfaces in the scene as polyhedra. Nodes represent primitive topological elements (faces, edges, vertices, objects, and edge groups) or primitive geometric elements (planes, lines and points). Two types of links are used - part-of link (relation between two topological nodes) and the geometric constraint link (representing the constraint relation between a geometric and topological node). Generates a surface-based 3-D scene model from the wire frame descriptions utilizing heuristics of building shapes.	Domain knowledge used is viewpoint dependent. Can generate 3-D perspective views of reconstructed buildings from any desired viewpoint. Uses truth maintenance procedures when hypotheses need modification.

Developer: Faugeras and Price (1980), USC
Purpose: Semantic Description of Aerial Images
Sample Domains: Search for known objects in an aerial image

Table III-1
Model-Based Vision Systems

Approach	Modeling and Representation	Remarks
Construct a network of image segments using region-based image segmentation and linear feature extraction. Solution is by stochastic matching (relaxation) of the image network to a portion of the model network. The approach takes the form of a constrained optimization graph search. Names are assigned to units with matching probabilities above 80%.	Represent image segments by properties: -average color -simple texture measure -position -orientation -simple shape measures Relations between image segments include adjacency, proximity and relative position Model is described by a semantic network, with the nodes being segments projected into the image plane. The arcs are positional relationships. Image is also represented by the same type of semantic network as model.	Uses image-dependent models restricted in viewpoint Segmentation is relatively weak.

APPENDIX J

Tables of

Commercially Available Vision Systems

Commercially Available Industrial Vision Systems
Table VI

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
G.E. Syracuse, NY	Optomation II Approx. $50K	This is a sophisticated high speed vision system designed primarily for inspection and measurement. (The system can do measurements along user defined lines.) Optomation uses up to 4 G.E. solid state CID cameras for input. The system is based on a multi-distributed microprocessor architecture partitioned to take maximum advantage of hardware, firmware and software modules to achieve high speed, flexibility and low cost. The system first thresholds the image to binary. The thresholded images are in a 256x256x8 bit high speed dynamic random access memory, normally operated as four 256x256 2 bit pages. The thresholded image is next "windowed" to establish spatial limits for data to be further processed. The windows are raster-scanned to find edge-points (where pixels change from 0 to 1 or vice versa). Using a patented corner point encoder the system observes where edges change direction (45° or greater) and labels these"cornerpoints." Only the corner points are stored. This is all accomplished in a single pass. The stored corner points are then correctly associated such that each object or item (closed edge set) is reconstructed and stored in an item file. The feature extracter then analyzes these item files and extracts key features such as an area, centroid, bounding rectangle, distances, angles, etc. (similar to the SRI module). Approximately 50,000 corner points per second can be processed. The processor can simultaneously operate on4 scenes, composed of 64 objects with up to 3300 corner points. The system can thus handle up to 15 images per second for each of 4 asynchronous cameras. Optomation II can be readily programmed by the user in Basic-like VPL (Vision Planning Language).	X	X		

Commercially Available Industrial Vision Systems
Table VI (continued)

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
Machine Intelligence Corp. Sunnyvale, CA	VS100 Vision System Approx. $35K	This sophisticated system is based on the SRI Vision Module System and is easily programmed using a light-pen controlled menu on TV monitor. Can also be programmed in BASIC on a DS-100 Development System. VS100P, aportable version, is also available.		X	X	X
	VS110 Vision System	This system adds a Programmable Image Overlay-feature to the VS100, so that by masking, or differencing,a precisely located part can be inspected for flaws or tolerance verification.	X	X	X	X
	UNIVISION	This system is basically the VS100 used as a pattern recognition system to provide a vision sensing capability for UNIMATE PUMA robots. It consists of a vision processor, graphic display monitor, light pen, camera(s) and UNIMATION's VAL software and hardware interface. The system is designed to operate in real time. UNIVISION can be trained to recognize a maximum of 9 different objects with up to 12 non-occluded parts in the scene at one time. For each part, 13 distinguishing features can be generated including area, perimeter, C.G., number of holes, and maximum and minimum radii.	X	X	X	X
	DS-100 Development System Approx. $95K	Allows easy vision program development by engineers with little computer experience. Shields the user from most computer-related details while providing sophisticated development tools, such as 20 megabytes of disk, a file system, several screen-oriented editors, a compiler, and several debugging aids. Programs developed on the DS-100 can be executed in the factory on a VS-100 or VS-110. Price includes an integral vision system, training, and user-support documentation.				

Commercially Available Industrial Vision Systems
Table VI (continued)

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
Automatix, Inc. Burlington, Mass	Autovision II Approx. $35K	This sophisticated system has many of the aspects of the SRI Vision Module. It can also window images for template matching or feature extraction. System has frame buffer storage and can handle sixteen gray levels. System is user programmable in a customized high level robotics system language called RAIL. The system is either obtainable as a stand-alone, or is available as an option to Automatix robots for assembly and arc welding (for seam-tracking guidance utilizing structured light).	X	X	X	X

Commercially Available Industrial Vision Systems
Table VI (continued)

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
Octek, Inc. Burlington, Mass.	Robot Vision Module 2200 Approx. $9.9K (without camera or primary computer)	This is a sophisticated computer system designed to be interfaced to a DEC or Data General computer. 50 FORTRAN S/W programs come with it to provide the capability of the SRI vision Module. Can inspect up to 5 parts/sec. using a modified SRI algorithm and feature vectors (having components such as area, moments, etc.). Can handle up to 50 objects in a scene at once. System incorporates a frame-grabber which can handle 4 images at once. System can also do signal averaging and "kernal manipulation" (spatial filtering, template matching or image subtraction). System can do histograms, measure objects in terms of length, width and angle. Can also do psuedo color with gray scale. Octek is now supplying a HITACHI 320x240 resolution solid-state miniature 4-bit gray level camera. Octek also supplies CCD and other solid-state cameras, as well as monochrome and RGB monitors.	X	X	X	X
	20/20 Vision Development System Approx. $45K	An integrated self-contained package containing a camera, Image Processor, 11/23 computer, B & W monitor, printer and cabinet, plus FORTRAN subroutines for users to implement their own applications.				

Commercially Available Industrial Vision Systems
Table VI (continued)

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
Object Recognition Systems New York, N.Y.		Computer-based vision systems that uses pattern recognition techniques for high-speed verification, packaged for industrial use. The standard products have "on-the-fly" hardware (no frame-grabber). Systems sample and filter with analog signals. Using firmware generated windows, can zoom in on particular areas. General approach is to extract gray scale samples, then extract features and compare with stored patterns. Also available is a picture differencing algorithm (with average of previous frames) for change detection. Systems employ gray scale, edge detection, and texture information, as appropriate.	X	X	X	X
	System Q Approx. $20K	For alpha-numeric legibility verification. Does edge of characters detection (proprietary) on fly. Matches against stored prototype (constructed of samples).	X			
	ScanSystem 100 Approx. $20K	Real time vision system for verification and inspection - single pattern library - 300 images per minute.	X	X		
	ScanSystem 200 Approx. $25K	Real time vision system for verification, inspection and recognition. 160 Pattern library - 300 images per minute.	X	X	X	
	ScanSystem 1000 Approx. $65K	Designed for keyboard verification. Train it with a good keyboard under joystick control to bring each key into view of an area-type scanner. Forms a file of table addresses, windows and associated features. Can inspect keyboards at rate of one/minute. Can also inspect populated PCB boards for correctness of component placement. Checks for color (via gray scale) and height and width of characters. Checks for maximum correlation and extracts transform-coded edge features and matches them using a pattern distance measure. Also generates statistical quality control information.	X	X		
	i-bot Vision System Approx. $30K	Vision system designed specifically to assist robots to remove individual objects from a jumbled bin of parts. Module can guide the pickup of jumbled cylindrical and spherical shaped objects from a bin , using a modification of the U. of Rhode Island peak reflection technique. Development is underway on bin-picking for a greater variety of shapes. A 3-D vision system using photo-metric stereo is also under development.		X	X	X

Commercially Available Industrial Vision Systems
Table VI (continued)

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
Spectron Engr. Inc. Denver, CO		Used primarily for dimensional measurement and defect detection and evaluation. Detector based on photo diode arrays. Uses cameras, processors and controllers of their own design. Front end optics, light sources and software tend to be application specific. Their strength is in high resolution applications. Have a library of subroutines to draw upon in devising custom applications.				
	CE 400/410 Wire, Optical Fiber Diameter Measurement System Approx. $10K	Uses a flash source to overcome vibration. Measurements at 30 to 60/sec. Runs automatically. Can be customized for process control.		X		
	CE 400/410 Print Photocopier Scanning Microdensitometer	Looks for print sharpness and uniformity in evaluating photocopiers. Can also evaluate paperstock. Differences adjacent elements in letters to obtain a mean square difference in reflectivity.		X		
	Customized Dimensional Measurement Systems Approx. $20K to $35K	First digitally corrects data for distortion, photodiode variations etc. using 256 gray levels. Uses image reconstruction techniques to enhance resolution.	X	X	X	

Commercially Available Industrial Vision Systems
Table VI (continued)

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
Videometrix Chatsworth, CA	VPU (Video Processor Unit) Approx. $15K Approx. $20K with camera and monitor	Converts a video analog TV camera output to a binary 500x400 image. Basic capability of Vision Module is programmed in firmware (PROM), enabling the system to calculate features such as c.g., area, edges, diameter, etc., in 50 milliseconds. System can operate in a stand-alone mode or interface to a computer.	X	X		
	VM-2000 Measurement System Approx. $70K for a completely automated measurement station	System can window and automatically focus in on a surface point so that it can measure to 0.0001" in a 4" cubic region. Using a 40 power microscope, measurements to 7 millionths of an inch are attainable.				

Commercially Available Industrial Vision Systems
Table VI (continued)

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
Copperweld Robotics Troy, Michigan	Opto-Sense Approx $40K and up	Uses a multiple windowing technique, setting up subsets of rectangles around portions of interest such as holes. Sets threshold limits and does area counts above or below threshold in a window to see if portion is within tolerance. Requires part be properly oriented.	X	X		
		Can be upgraded to incorporate SRI Vision Module features	X	X	X	
		Can be further upgraded with a full-frame grabber (up to 256 levels of gray) and customized software added, for more sophisticated applications.	X	X	X	X
Robot Vision Systems Mellville, NY	Primarily Custom Vision Systems for Military and Industry	A company strong point is propriatary techniques in the use of structured light and triangulation to discern the 3D coordinates of an object under view. This volumetric vision ("solid photography") approach can capture an enormous amount of data very quickly (all the data required to define a man's head in 0.9 seconds). They also utilize area type sensors for robot vision. Have made a sensor system for Cummins Diesel to measure very large unfixtured engine block castings. The system makes 1250 measurements in 35 minutes and compares dimensions with those stored in a computer. The system is designed to achieve accuracy of 0.0001". Now building a robotic welding system for use with a Millicron T3 robot to weld automobile frames. This is a two part system where first the weld line is scanned at 180"/sec. to determine seam locations and width. Using this information, the seam is then adaptively welded. Now making a standard vision system with a 1" or 4" field of view with an accuracy of 0.001" for inspection applications, but which could also be configured for welding.	X	X	X	X

Commercially Available Industrial Vision Systems
Table VI (continued)

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
Cognex Boston, Mass.	Dataman Approx. $25K for basic OCR system.	System uses a DEC PDP-11-23 and other off-the-shelf hardware such as cameras. Dataman derived from research Dr. Shillman did at M.I.T. on how humans recognize patterns. Implemented resulting proprietary algorithms can read badly degraded alpha-numerics. System can read virtually any alpha-numerics humans can (will reject unreadables rather than make errors). Basic system is for optical code reading (font specific). System can also be used for print quality assurance (legibility) and quality control.	X	X	X	

Commercially Available Industrial Vision Systems
Table VI (continued.)

Company	System	Comments	Verification	Inspection	Recognition	Manipulation
Ham Industries Macedonia, OH	Ham Scan 3000 Approx. $6K (without camera and monitor)	This is a verification and inspection system that operates on an analog image either by template matching or by analog integration. Can do windowing, or measurements on a single programmed line. Windowing, thresholding, line placements, etc. is set using factory customized software in the integral micro-processor. System can be trained by showing, or by manual use of switches.	X	X		
	Ham Scan 1000 Approx. $4K	Does template matching with adjustable allowable deviationsusing gray scale and a single window.	X	X		
	Ham Scan 2000	Similar to the 1000, except that it has a double window.	X	X		

APPENDIX K

GLOSSARY

APPENDIX K

GLOSSARY *

- Artificial Intelligence (AI) Approach: An approach that has its emphasis on symbolic processes for representing and manipulating knowledge in a problem solving mode.

- Bar Operators: Convolution masks to detect second derivatives of image brightness in particular directions.

- Binary Image: A black and white image represented as zeros and ones, in which objects appear as silhouettes.

- Blackboard Approach: A problem solving approach whereby the various system elements communicate with each other via a common working data storage called the blackboard.

- Blob: A connected region in a binary image.

- Blocks World: Scenes consisting of three dimensional polyhedral object configurations. A simple artificial world used to explore computer vision concepts.

- Bottom Up (Data Driven): Refers to the sequential processing by a vision system, beginning with the input image and terminating in an interpretation.

- CAD/CAM: Computer-aided design / computer-aided manufacture.

- Chain Code: A boundary representation which starts with an initial point and stores a chain of directions to successive points.

- Computer Vision (Computational or Machine Vision): Perception by a computer, based on visual sensory input, in which a concise description is developed of a scene depicted in an image. It is a knowledge-based, expectation-guided process that uses models to interpret sensory data. Used somewhat synonymously with image understanding and scene analysis.

- Concurve: A boundary representation consisting of a chain of straight lines and arcs.

- Convolve: Superimposing a mxn operator over a mxn pixel area (window) in the image, multiplying corresponding points together and summing the result.

- Corner: An abrupt change in direction of a curve.

- Correlation: A correspondence between attributes in an image and a reference image.

---

* As yet no standard definitions exist, so that the definitions listed here can be considered to be somewhat imprecise.

- Description: A symbolic representation of the relevant information, e.g., a list of statistical features of a region.

- Digitized Image: A representation of an image as an array of brightness values.

- Domain: The sphere of concern. The task world. A set of allowable inputs.

- Edge: A change in pixel values (exceeding some threshold) between two regions of relatively uniform values. Edges correspond to changes in brightness which can correspond to a discontinuity in surface orientation, surface reflectance or illumination.

- Edge Operators: Templates for finding edges in images.

- Edge-Based Stereo: A stereographic technique based on matching edges in two or more views of the same scene taken from different positions.

- Features: Simple image data attributes such as pixel amplitudes, edge point locations and textural descriptors, or somewhat more elaborate image patterns such as boundaries and regions.

- Feature Vector: A set of features of an object (such as area, number of holes, etc.) that can be used for its identification.

- Feature Extraction: Determining image features by applying feature detectors.

- Gaussian Filtering: A convolution proceedure in which the weighting of pixels in the template fall off with distance according to a Gaussian distribution.

- General Purpose Vision System: A vision system that is universally applicable. A system that is based on generic rather than specific knowledge (cf. Nevatia, 1982, p 188). A system that can deal with unfamiliar or unexpected input.

- Generalized Cone (Generalized Cylinder): A volumetric model defined by a space curve, called the spine or axis, and a planar cross section normal to the axis. A "sweeping rule" describes how the cross section changes along the axis.

- Generalized Ribbon (See Skeleton Representation): A planar region approximated by a medial line (axis) and the perpendicular distances to the boundary. The 2-D version of a generalized cone.

- Global Method: A method based on non-local aspects, e.g., region splitting by thresholding based on an image histogram.

- Goal Driven: Top-down approach.

- Gradient Space: A coordinate system (p,q) in which p and q are the rates of change in depth (gray value) of the surface of an object in the scene along the x and y directions (the coordinates in the image plane). Thus (p,q,1) has the direction of the surface normal.

- Gradient Vector: The orientation and magnitude of the rate of change in intensity at a point in the image.

- Graph (Also Relational Graph): An image representation in which nodes represent regions and arcs between nodes represent properties of and relations between these regions.

- Gray Level: A quantized measurement of image irradiance (brightness), or other pixel property.

- Heterarchical Approach: An image interpretation control structure in which no processing stage is in sole command, but in which each stage can control other stages to its needs as required.

- Heuristics: "Rules of thumb," knowledge or other techniques used to help guide a problem solution.

- Hierarchical Approach: An approach to vision based on a series of ordered processing levels in which the degree of abstraction increases as we proceed from the image level to the interpretation level.

- Higher Levels: The interpretative processing stages such as those involving object recognition and scene description, as opposed to the lower levels corresponding to the image and descriptive stages.

- Histogram: Frequency counts of the occurrence of each intensity (gray level) in an image.

- Hough Transform: A global parallel method for finding straight or curved lines, in which all points on a particular curve map into a single location in the transform space.

- Hueckel Operator: A method for finding edges in an image by fitting an intensity surface to the neighborhood of each pixel and selecting surface gradients above a chosen threshold value.

- Iconic: Image-like.

- Image: A projection of a scene into a plane. Usually represented as an array of brightness values.

- Image Processing: Transformation of an input image into an output image with more desirable properties, such as increased sharpness, less noise, and reduced geometric distortion. Signal processing is a 1-D analog.

- Image Understanding (IU): Employs geometric modeling and the AI techniques of knowledge representation and cognitive processing to develop scene interpretations from image data. IU has dealt extensively with 3D objects. IU usually operates not on an image but on a symbolic representation of it. IU is somewhat synonymous with computer vision and scene analysis.

- Irradiance: The brightness of a point in the scene.

- Isomorphic Representation: A representation in which there is a one to one correspondence between the scene and its representation, (e.g., an image or a map).

- Interpretation: Establishing a correspondence between the scene and a set of models. Assigning names to objects in a scene.

- Interpretation-Guided Segmentation: Using models to help guide image segmentation, by the process of extending partial matches.

- Intrinsic Characteristics: Properties inherent to the object, such as surface reflectance, orientation, incident illumination and range.

- Intrinsic Images: A set of arrays in registration with the image array. Each array corresponds to a particular intrinsic characteristic.

- Laplacian Operator: The sum of the second derivatives of the image intensity in the x and y directions is called the Laplacian. The Laplacian operator is used to find edge elements by finding points where the Laplacian is zero.

- Line: A thin connected set of points contrasting with neighbors on both sides. Line representations are extracted from edges.

- Line Detectors: Oriented operators for finding lines in an image.

- Line Followers: Techniques for extending lines currently being tracked.

- Low Level Features: Pixel-based features such as texture, regions, edges, lines, corners, etc.

- Model-based Vision System: A system that utilizes a priori models to derive a desired description of the original scene from an image.

- Module: A processing unit in a vision system.

- Monocular: Pertaining to an image taken from a single viewpoint.

- Optical Flow: The distribution of velocities of apparent movement in an image caused by smoothly changing brightness patterns.

- Pattern Recognition: A technique that classifies images into predetermined categories, usually using statistical methods.

- Perception: An active process in which hypotheses are formed about the nature of the environment, or sensory information is sought to confirm or refute hypotheses.

- Photometric Stereo: An approach in which the light source illuminating the scene is moved to different known locations, and the orientation of the surfaces deduced from the resulting intensity variations.

- Pixel (Picture Element): The individual elements in a digitized image array.

- Primal Sketch: A primitive description of the intensity changes in an image. It can be represented by a set of short line segments separating regions of different brightnesses.

- Pyramid: A hierarchical data structure that represents an image at several levels of resolution simultaneously.

- Quadtree: A representation obtained by recursively splitting an image into quadrants, until all pixels in a quadrant are uniform with respect to some feature (such as gray level).

- Recognition: A match between a description derived from an image and a description obtained from a stored model.

- Reflectance (Albedo): The ratio of total reflected to total incident illumination at each point.

- Region: A set of connected pixels that show a common property such as average gray level, color or texture, in an image.

- Region Growing: Process of initially partitioning an image into elementary regions with a common property (such as gray level) and then successively merging adjacent regions having sufficiently small differences in the selected property, until only regions with large differences between them remain.

- Registration: Processing images to correct geometrical and intensity distortions, relative translational and rotational shifts, and magnification differences between one image and another or between an image and a reference map. When registered, there is a one to one correspondence between a set of points in the image and in the reference.

- Relaxation Approach: An iterative problem solving approach in which initial conditions are propagated utilizing constraints until all goal conditions are adequately satisified.

- Relational Graph: See "graph."

- Representation: A symbolic description or model of objects in the image or scene domain.

- Run-length Encoding: A data compression technique in which an image is raster-scanned and only the lengths of runs of consecutive pixels with the same property are stored.

- Scene: The 3-D environment from which the image is generated.

- Scene Analysis: The process of seeking information about a 3-D scene from information derived from a 2-D image. It usually involves the transformation of simple features into abstract descriptions.

- Segmentation: The process of breaking up an image into regions (each with uniform attributes) usually corresponding to surfaces of objects or entities in the scene.

- Semantic Interpretation: Producing an application-dependent scene description from a feature set (representation) derived from the image.

- Semantic Network: A representation of objects and relationships between objects as a graph structure of nodes and labeled arcs. See "graph."

- Skeleton Representation (See "Generalized Ribbons"): A representation of a 2-D region by the medial line and the perpendicular distance to the boundary at each point along it.

- Sketch Map: A rough line drawing of a scene.

- Sobel Operator: A popular convolution operator for detecting edges. Similar to other difference operators such as the Prewitt Operator.

- Spectral Analysis: Interpreting image points in terms of their response to various light frequencies (colors).

- Splines (B-Splines): Piecewise continuous polynomial curves used to approximate a curve.

- Steroscopic Approach: Use of triangulation between two or more views, obtained from different positions, to determine range or depth.

- Structured Light: Sheets of light and other projective light configurations used to directly determine shape and/or range from the observed configuration that the projected line, circle, grid, etc. makes as it intersects the object.

- Symbolic Description: Non-iconic scene descriptions such as graph representations.

- Syntactic Analysis: Recognizing images by a "parsing" process as being built up of primitive elements.

- Template: A prototype iconic model that can be used directly to match to image characteristics for object recognition or inspection.

- Template Matching: Corellating an object template with an observed image field - usually performed at the pixel level.

- Texture: A local variation in pixel values that repeats in a regular or random way across a portion of an image or object.

- Thresholding: Separating regions of an image based on pixel values above or below a chosen (threshold) value.

- Top Down Approach (Goal Directed): An approach in which the interpretation stage is guided in its analysis by trial or test descriptions of a scene. Sometimes referred to as "Hypothesize and Test."

- Tracking: Processing sequences of images in real time to derive a description of the motion of one or more objects in a scene.

- Vertex: The point on a polyhedron common to three or more sides.

- Viewpoint: The position (or direction) from which the scene is observed.

- Vision: The process of understanding the environment based on image data.

- Wireframe Model: A 3-D model, similar to a wireframe, in which the object is defined in terms of edges and vertices.

- Window: A selected portion (usually square or rectangular) of an image.

- 2-D: Two dimensional.

- 2.5-D Sketch: A scene representation proposed by Marr (1978), consisting of surface distances and orientations.

- 3-D: Three dimensional.

APPENDIX L

SOME PUBLICATION SOURCES FOR FUTHER INFORMATION

APPENDIX L

SOME PUBLICATION SOURCES FOR FURTHER INFORMATION

A. Recent Books

Ballard, D.H., and Brown, C.M., Computer Vision, Englewood Cliffs: Prentice Hall, 1982.

Brady, M. (Ed.) Computer Vision, Amsterdam: North Holland, 1981.

Cohen, P.R., and Feigenbaum, E.A., "Vision," The Handbook of Artificial Intelligence, Vol. III, Los Altos, CA: Kaufman, 1982, pp. 125-321.

Haralick, R. (Ed.) Picture Data Analysis, Berlin: Springer-Verlag, 1982.

Marr, D.C., Vision, San Francisco: Watt. Freeman, 1982.

Nevatia, R., Machine Perception, Englewood Cliffs: Prentice Hall, 1982.

Rosenfeld, A., and Kak, A.C., Digital Image Processing, 2nd Ed., Vols. 1 and 2, New York: Acad. Pr., 1982.

Pavlidis, T., Algorithms for Graphics and Image Processing, Rockville, Md.: Computer Science Press, 1982.

Hord, R. M., Digital Image Processing of Remotely Sensed Data, New York: Acad. Pr., 1982

B. Periodic Conference and Workshop Proceedings

DARPA Image Understanding Workshops - Sci. Applications Inc.

International Conferences on Pattern Recognition - IEEE Computer Society

National Conferences on Artificial Intelligence - AAAI

IEEE Workshops on Computer Vision - IEEE Computer Society

International Joint Conferences on Artificial Intelligence

International Conferences on Robot Vision - The Industrial Robot Journal and Sensor Review

Workshops on Industrial Applications of Computer Vision - IEEE Computer Society

SPIE Technical Symposia - Society of Photo-Optical Instrumentation Engineers

NSF Workshops (Aperiodic workshops on various topics in computer vision).

C. Periodicals

Computer Graphics and Image Processing

Artificial Intelligence

IEEE Transactions on Pattern Analysis and Machine Intelligence

Pattern Recognition

International Journal of Robotics Research

IEEE Transactions on Systems, Man and Cybernetics

D. Some Recent Bibliographies, Surveys and Syntheses

Ahuja, N. and Schacter, B., "Image Models," ACM Computing Surveys, Vol. 13, No. 4, Dec. 1981, pp. 373-398.

Barrow, H.G. and Tenenbaum, J. M., "Computational Vision," Proc. of the IEEE, Vol. 69, No. 5, May 1981, pp. 572-595.

Binford, T.O., "Survey of Model-Based Image Analysis Systems," Robotics Research, Vol. 1, No. 1, Spring 1982, pp 18-64.

Brady, M., "Computational Approaches to Image Understanding," Computer Surveys, Vol. 14, No. 1, March 1982, pp. 3-71.

Chin, R. T., "Automated Visual Inspection Techniques and Applications: A Bibliography," Pattern Recongition, Vol. 5, No. 4., 1982, pp. 328-357.

Genery, D., Cunningham, R., Saund, E., High, J., and Rouff, C., Computer Vision, JPL 81-92, JPL, Pasadena, CA, Nov. 1, 1981.

Kruger, R.P. and Thompson, W.B., "A Technical and Economic Assessment of Computer Vision for Inspection and Robotic Assembly," Proc. of the IEEE, Vol. 69, No. 12, Dec. 1981, pp. 1524-1538.

Rosenfeld, A., Picture Processing, U. of Md. C.S. Center, College Park, Md., A yearly bibliography of computer processing of pictorial information.

Srihari, S. N. "Representation of 3 D Digital Images," ACM Computing Surveys, Vol. 13, No. 4 Dec. 1981., pp. 399-424.

NBS-114A (REV. 2-8C)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET (See instructions)	1. PUBLICATION OR REPORT NO. NBSIR82	2. Performing Organ. Report No.	3. Publication Date September 1982

4. TITLE AND SUBTITLE

AN OVERVIEW OF COMPUTER VISION

5. AUTHOR(S)

William B. Gevarter

6. PERFORMING ORGANIZATION (If joint or other than NBS, see instructions)

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

7. Contract/Grant No.

8. Type of Report & Period Covered

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (Street, City, State, ZIP)

10. SUPPLEMENTARY NOTES

☐ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

11. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here)

This report provides on overview of computer vision. The emphasis is on image understanding and scene analysis, though pertinent aspects of pattern recognition are treated. Image processing for sensor correction, rectification, image enhancement, etc., is not covered.

This report reviews the basic approach to computer vision systems, the techniques utilized, applications, the current existing systems and state-of-the-art, issues and research requirements, who is doing it and who is funding it, and finally future trends and expectations. The intent is to provide an overall perspective of this vital field with its many participants, that will be useful to engineering and research managers, potential users and others who will be impacted by this field as it unfolds.

12. KEY WORDS (Six to twelve entries; alphabetical order; capitalize only proper names; and separate key words by semicolons)

Artificial intelligence; Automation; Computational; Computer perception; Computer vision: Forecasting; Industrial vision systems; Image understanding; Pattern recognition; Scene analysis; Vision; Vision systems

13. AVAILABILITY

☒ Unlimited

☐ For Official Distribution. Do Not Release to NTIS

☐ Order From Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

☒ Order From National Technical Information Service (NTIS), Springfield, VA. 22161

14. NO. OF PRINTED PAGES 170

15. Price $15.00